Volume 17
Number 12
25 March 2026
Pages 5763–6244

Chemical Science

rsc.li/chemical-science

ISSN 2041-6539

ROYAL SOCIETY OF CHEMISTRY

PERSPECTIVE
Hao Li *et al.*
Digital materials ecosystem: from databases to AI agents for autonomous discovery

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

PERSPECTIVE

View Article Online
View Journal | View Issue

Cite this: *Chem. Sci.*, 2026, **17**, 5782

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

Digital materials ecosystem: from databases to AI agents for autonomous discovery

Di Zhang, Xue Jia, Yuhang Wang, Heng Liu, Qian Wang, Seong-Hoon Jang, Daksh Shah, Songbo Ye, Hung Ba Tran and Hao Li *

Received 25th November 2025
Accepted 20th February 2026

DOI: 10.1039/d5sc09229a

rsc.li/chemical-science

The concept of a digital materials ecosystem represents a new paradigm in materials research, where data, theory, and automation are integrated into a unified and iterative framework. By combining reliable databases, physical frameworks, and intelligent data analysis, materials discovery is evolving from empirical exploration toward a systematic and predictive science. The rapid growth of data and artificial intelligence (AI) has enabled the identification of complex structure–property relationships, while advances in automated synthesis and high-throughput characterization are closing the loop between prediction and validation. Looking forward, the field must focus on building trustworthy and benchmarked datasets, developing interpretable and high-precision models, and designing AI tools that embody human scientific reasoning. Equally important is ensuring standardization and consistency between digital inputs and experimental responses. Together, these efforts will transform materials discovery from data accumulation into genuine knowledge generation, paving the way for an autonomous and self-improving research ecosystem that accelerates both fundamental understanding and technological innovation.

1. Introduction

The digital materials ecosystem represents a transformative paradigm in materials research, where the integration of data, theory, and automation is reshaping how new materials are discovered, optimized, and applied (Fig. 1). Traditionally, materials science relied heavily on empirical trial-and-error approaches to discover and develop new materials. However, with the rapid growth of data generation, artificial intelligence (AI), and automation technologies, new approaches have emerged that leverage vast datasets, advanced computational tools, and high-throughput experimental techniques to accelerate the pace of materials discovery and innovation.[1]

At the heart of this ecosystem are materials databases, which serve as the backbone for aggregating experimental and theoretical data.[2] These large-scale databases allow for the efficient retrieval, analysis, and reuse of information across diverse materials systems, providing the foundation for subsequent data-driven research. As materials databases grow in size and complexity, they enable deeper insights into structure–property relationships, fostering a more systematic and predictive approach to material design.

In parallel, machine learning (ML) and AI-driven modeling are advancing the capabilities of materials science.[3,4] These technologies enhance predictive accuracy by learning from historical data, identifying complex patterns that may be difficult for traditional methods to uncover. AI models can now predict material properties, suggest new material candidates, and even guide experimental design, all of which are pivotal in reducing the time and cost involved in material development. Moreover, the integration of automated synthesis and high-throughput characterization techniques has led to the development of a closed feedback loop in materials research.[5,6] In this loop, predictions made by AI models are validated through experiments, and new data generated from experiments are fed back into the system, continuously refining the models. This self-evolving cycle fosters a more efficient and dynamic approach to materials discovery, where the pace of innovation is

Advanced Institute for Materials Research (WPI-AIMR), Tohoku University, Sendai 980-8577, Japan. E-mail: li.hao.b8@tohoku.ac.jp

Hao Li

Hao Li is a Distinguished Professor at the Advanced Institute for Materials Research (WPI-AIMR), Tohoku University, Japan. His research focuses on developing AI, materials theory, and autonomous experimentation for closed-loop materials design. He proposed the concept of the digital materials ecosystem and serves as the founding Editor-in-Chief of the journal AI Agent.

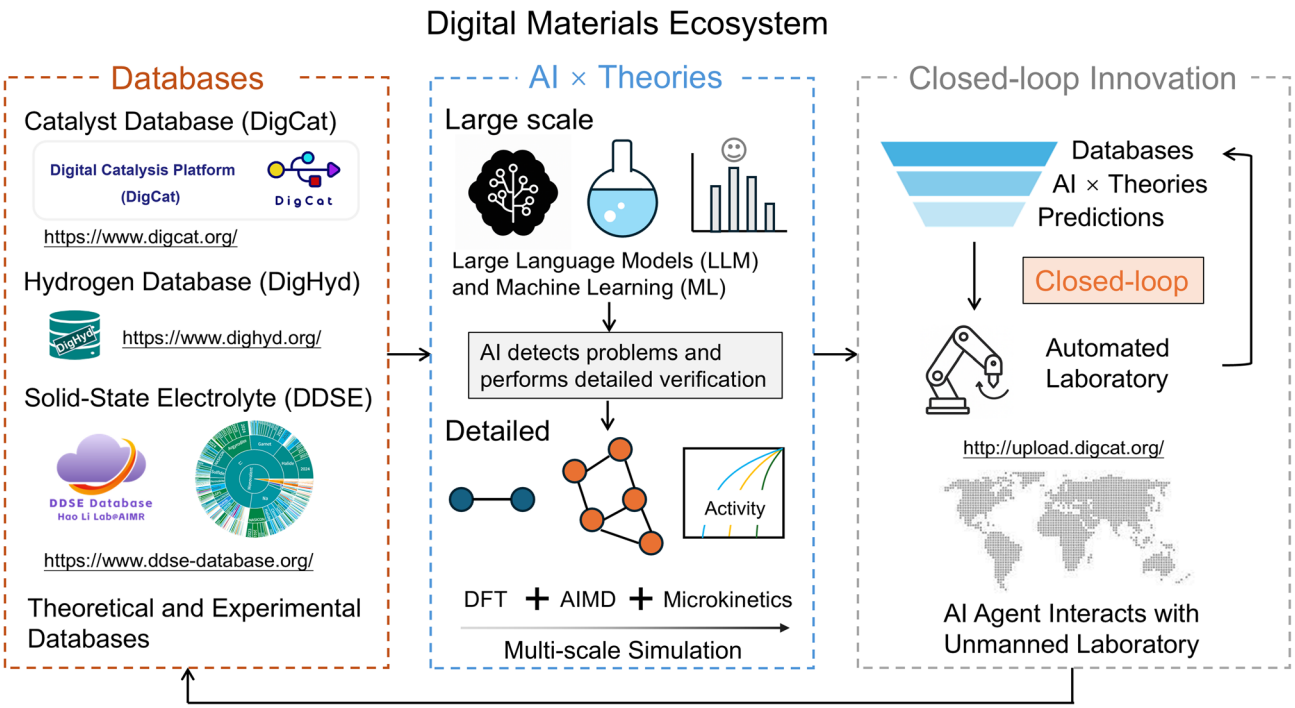

Fig. 1 Schematic illustration of the concept of the digital materials ecosystem. This concept consists of databases (left frame), AI and theoretical frameworks (middle frame), and closed-loop innovation based upon modern experimental techniques (right frame).

accelerated, and previously unattainable breakthroughs are within reach.

The digital materials ecosystem is not limited to specific material classes. It spans a wide range of domains, from solid-state batteries and catalysis to hydrogen storage and beyond. Each of these material systems contributes uniquely to the broader ecosystem, where AI, automation, and data-driven methodologies are applied to solve complex problems. By drawing on examples from various material classes, this perspective highlights the versatility and wide applicability of the digital materials ecosystem, emphasizing its potential to revolutionize materials research across diverse application areas.

2. The foundation of modern digital materials: databases and data infrastructure

Researchers across various fields have increasingly emphasized the development of databases in recent years, with efforts spanning multiple levels: from specialized small-domain databases targeting specific materials or applications,[7–11] to experimental data-driven databases,[12–14] and computational databases based on high-throughput calculations and theoretical predictions.[15–24] These advancements have supported the growth of a data-driven research ecosystem that connects fundamental research with industrial applications. As an example, we focus on catalysis, where databases and computational tools are transforming the discovery of new catalysts. By integrating machine learning with these databases, the process of catalyst discovery is being significantly accelerated.

2.1. Examples of small-domain databases

In 2020, Marchenko *et al.* developed an open database for the field of 2D hybrid perovskites, combining experimental data with computational and ML predictions.[7] The database includes crystallographic information (CIF files, space groups, number of layers, structural types, bond lengths, bond angles, penetration depth, *etc.*), experimental band gaps, ML-predicted band gaps, and atomic partial charges. Its purpose is to reveal composition–structure–property relationships (QSPR) and provide tools for the rational design and high-throughput screening of new 2D hybrid perovskites. Similarly, in 2020, Sarkisov *et al.* created a database and toolchain for materials informatics, focusing on a computationally derived subset of MOFs from the Cambridge Structural Database (CSD).[8] This database, validated against Zeo++ and RASPA, includes geometric and accessibility characterization for approximately 12 000 MOFs, such as specific surface area, pore volume, pore size and limiting pore diameter, pore size distribution, connectivity dimensions, and density. The goal is to provide reproducible geometric characterization methods, improve consistency and comparability across algorithms, and enable structure–property relationship analysis and high-throughput screening using visualization and principal component analysis tools, thereby accelerating the discovery of new materials.

The strength of small-domain databases lies in their high specificity and focus, enabling the extraction of unique patterns within a given field and offering precise support for targeted research. However, their limitations include data fragmentation and a lack of compatibility and generalizability, making them

less suitable for cross-disciplinary studies and limiting their broader applicability.

2.2. Examples of experimental-data-driven databases

In 2002, Belsky *et al.* developed and refined the Inorganic Crystal Structure Database (ICSD), which contains over 60 000 entries of inorganic crystal structures.[13] The database includes chemical formulae, unit cell parameters, space groups, atomic coordinates, site occupancies, thermal parameters, and bibliographic information. It also provides a Windows graphical interface with search and visualization tools, offering reliable structural data for materials research, phase identification, and structure–property relationship analysis. In 2020, Huang and Cole *et al.* created a battery materials database by extracting experimental data from 229 000 research papers using ChemDataExtractor.[25] The database contains 17 354 chemical compounds and 292 313 entries, including capacity, voltage, conductivity, coulombic efficiency, energy, and their respective units and conditions. It provides a graphical user interface along with tools for data cleaning, standardization, and augmentation, enabling large-scale machine-readable data to support battery material design and prediction. In 2022, Ward *et al.* proposed the "Battery Data Genome (BDG)" database and data hub system, based on experimental multi-source data.[12] This framework spans data from materials and electrodes to single cells, modules/packs, and real-world systems, accompanied by complete metadata and standardized protocols. It aims to enable cross-stage data sharing and machine learning (ML) applications through unified standards and interoperable open-source software, accelerating battery material discovery, manufacturing optimization, and lifetime prediction, thereby facilitating efficient translation from research to deployment.

The strength of experimental data-driven databases lies in their high reliability, as they are derived from real experiments, providing critical parameters such as catalytic performance and battery properties. However, their limitations include slow data updates, limited coverage, and high experimental costs, which hinder the rapid generation of large-scale datasets and restrict their applicability in high-throughput screening and large-scale materials design.

2.3. Examples of computational databases

In 2013, Jain *et al.* developed the Materials Project database, based on computational data generated through high-throughput first-principles calculations.[16] The database provides open-access data on the crystal structures, energetics, and electronic structures of inorganic materials, supplemented by APIs and open-source analysis tools (*e.g.*, pymatgen). It enables rapid data retrieval, mining, and *in silico* "rapid prototyping" to accelerate materials innovation. In 2015, Qu *et al.* created the "Electrolyte Genome" database/platform for electrolytes, leveraging high-throughput quantum chemistry and automated workflows (*e.g.*, density functional theory, DFT).[15] The platform contains properties of 4830 molecules, including ionization potential/electron affinity, solvation, and ion pair dissociation (~55 000 calculations), along with tools for error correction, deduplication, and complex salt coordination generation. It facilitates large-scale screening and data-driven electrolyte molecular design.

The same year, Kirklin *et al.* developed the Open Quantum Materials Database (OQMD), conducting ~300 000 DFT calculations on ICSD structures and common prototypes.[20] The database includes crystal structures, total energies, formation energies, and chemical potential corrections, validated through large-scale comparisons with experiments (MAE $\approx$ 0.096 eV per atom). It enables thermodynamic stability assessments and predicts ~3200 potential new compounds, advancing materials discovery and design. In 2019, Winther *et al.* launched the Catalysis-Hub database, which aggregates over 100 000 adsorption/reaction energies and activation energies from DFT calculations.[23] It includes atomic structures, computational parameters, and APIs/web-based search tools. The platform supports reproducible, machine-readable data sharing and efficient screening, aiding the discovery and modeling of catalyst materials for sustainable energy applications. In 2022, Hu *et al.* introduced MaterialsAtlas.org, a materials informatics platform, and database integrating tools for composition/structure validation (*e.g.*, electroneutrality, Pauling rules, and dynamic stability), property predictions (*e.g.*, bandgap, elasticity, hardness, and thermal conductivity), and hypothetical material generation (*e.g.*, generated compositions and cubic structures).[22] The platform enables high-throughput exploration, screening, and visualization of inorganic crystals, significantly improving the efficiency of materials discovery and design. The Novel Materials Discovery (NOMAD) Laboratory provides one of the largest open repositories of computed materials data worldwide,[26] built around strict FAIR principles (findable, accessible, interoperable, and reusable). By aggregating and homogenizing raw first-principles calculations from multiple codes and users, NOMAD converts heterogeneous simulation outputs into a consistent, queryable data infrastructure that can be directly used for large-scale screening and data-driven model development. On top of this repository, the NOMAD Artificial Intelligence Toolkit offers workflows for feature extraction, dimensionality reduction, clustering, and supervised ML, enabling researchers to discover hidden patterns in high-dimensional materials spaces and to derive interpretable structure–property relationships. Other high-throughput frameworks such as AFLOW[27] further exemplify this transition from individual calculations to curated digital infrastructures. AFLOW provides an automated pipeline for generating, standardizing, and storing large numbers of first-principles calculations, together with symmetry analysis and a rich catalogue of derived materials properties accessible through the AFLOWLIB repository and programmatic APIs. In 2025, Huang *et al.* developed a comprehensive public single-atom catalyst (SAC) database and combined DFT-derived descriptors with ML models to rapidly screen 4d single-atom catalysts, identifying Rh_1B_4 and $Rh_1C_2S_2$ as highly active candidates for NO and Hg^0 oxidation.[28]

The advantages of computational databases lie in their ability to rapidly and efficiently generate theoretical predictions, making them suitable for large-scale material screening.

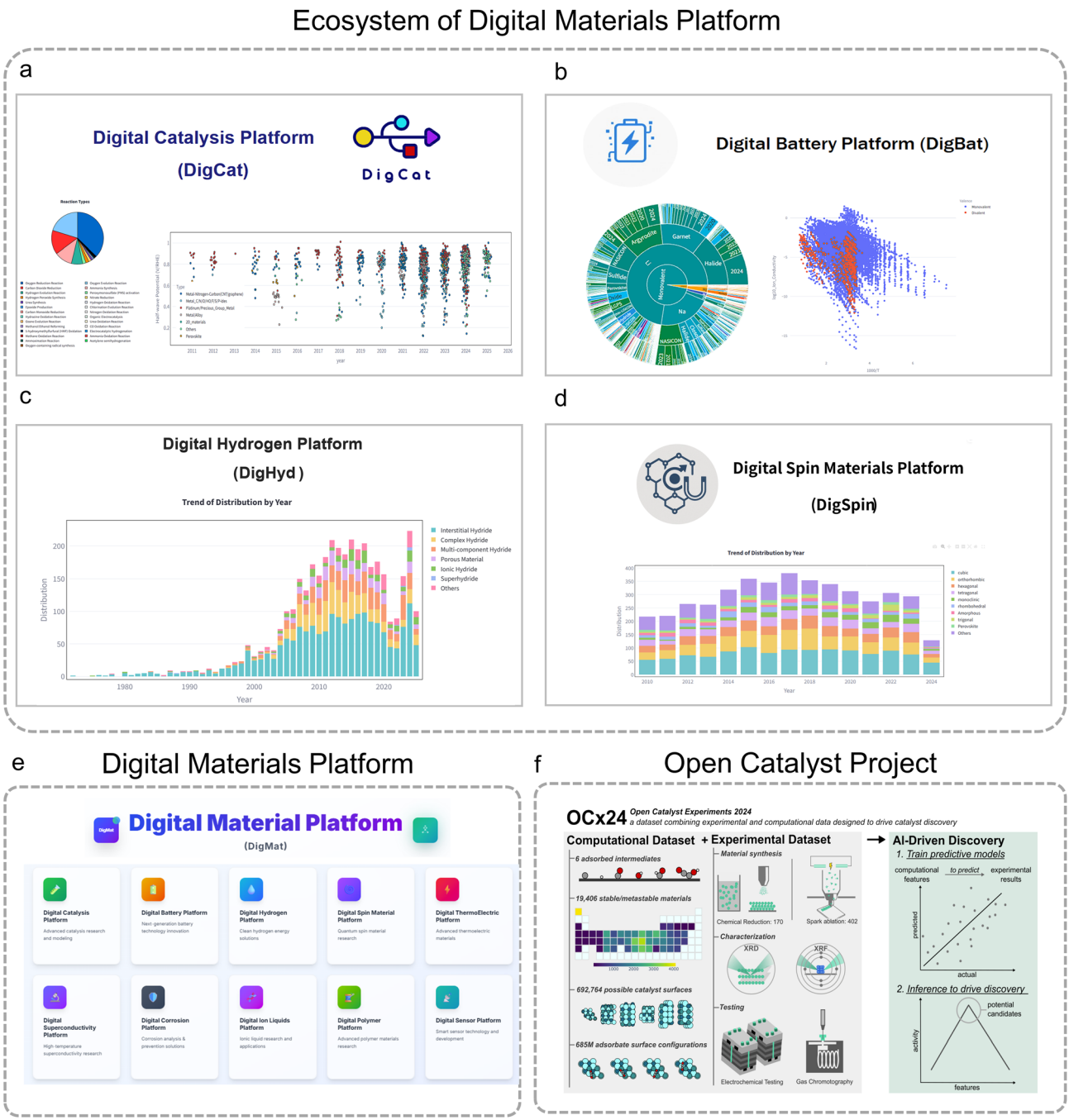

Fig. 2 Ecosystem of the Digital Materials Platform. (a) Digital Catalysis Platform (*DigCat*: https://www.digcat.org).[29] (b) Digital Battery Platform (*DigBat*: https://www.digbat.org).[30] (c) Digital Hydrogen Platform (*DigHyd*: https://www.dighyd.org).[31] (d) Digital Spin Materials Platform (*DigSpin*: https://www.digspin.org). (e) Digital Materials Platform (*DigMat*: https://www.digmat.org). (f) Open Catalyst Project.[32] Adapted with permission from: (f) ref. 32 © Copyright under a CC-BY 4.0 License.

They can also predict material properties that are difficult to measure experimentally, providing valuable insights. However, their limitations include potential inaccuracies in model predictions, which may not always align with rigorous experimental data. Consequently, computational data must be strictly experimentally validated before practical applications, limiting the direct applicability of such databases.

Encouragingly, there has been striking progress in the field in recent years. In 2024, Li and co-workers released the Digital Catalysis Platform (*DigCat*: https://www.digcat.org),[29] a catalysis database primarily based on experimental data combined with computational structures (Fig. 2a). *DigCat* encompasses over 400 000 experimental performance records and over 400 000 structural entries, enabling data visualization, literature tracking, AI-powered Q&A, cloud-based microkinetic simulations, and ML force field training, thereby accelerating catalysis research. In the same period, they launched the dynamic database of solid-state electrolytes (DDSEs) and the

Digital Battery Platform (*DigBat*: https://www.digbat.org) for solid-state electrolytes (SSEs) in solid-state batteries (Fig. 2b).[30] As of February 2026, DDSE contains data on over 3000 inorganic SSE materials, including ionic conductivities and activation energies measured across a wide temperature range (132.4–1261.6 K), covering diverse cationic and anionic systems. This database supports structure–property exploration and ML-based predictions. In 2026, the same team introduced the Digital Hydrogen Platform (*DigHyd*: https://www.dighyd.org, Fig. 2c),[31] a hydrogen storage materials database that integrates data from over 4000 publications (1972–2025) and more than 30 000 experimental data entries, including pressure–composition–temperature (PCT), temperature-programmed desorption (TPD), and discharge curves. This innovation drives data-driven discovery and significantly advances research in hydrogen storage materials. Furthermore, to accelerate the advancement of the digital materials ecosystem, the team has established a cutting-edge AI-powered digital platform for advanced materials discovery and development, termed the Digital Materials Platform (*DigMat*, https://www.digmat.org; Fig. 2e). This ecosystem encompasses a series of specialized sub-platforms, including the Digital Spin Materials Platform (*DigSpin*, Fig. 2d) for quantum spin and correlated materials, the Digital Thermoelectric Platform (*DigTEM*) for thermoelectric systems, the Digital Superconductivity Platform (*DigSuperC*) for superconductors, and the Digital Corrosion Platform (*DigCorrosion*) for materials corrosion analysis and prevention. Other related initiatives include the Digital Ionic Liquids Platform (*DigILS*), the Digital Polymer Platform (*DigPol*), the Digital Sensor Platform (*DigSen*), the Digital CO_2 Capture Platform (*DigCC*) and the Digital MOF Platform (*DigMOF*). Together, these dynamically updated platforms integrate millions of experimentally measured data and terabytes of literature-derived information, providing a robust foundation for the future expansion of the digital materials paradigm. Recently, the OCx24 (ref. 32) (Fig. 2f) study also provided high-throughput, AI-oriented experimental–computational datasets for electrocatalysis, connecting adsorption-energy descriptors with industrially relevant hydrogen evolution reaction (HER) and CO_2 reduction reaction (CO_2RR) performance. By revealing a data-driven Sabatier volcano and demonstrating transferable predictive capability across diverse material classes, OCx24 highlights how standardized, ML-ready experimental workflows can substantially narrow the gap between computation and practical catalyst discovery.

2.4. Limitations of existing databases and the urgent need for new development

The rapid advancements in energy and catalysis research have highlighted persistent challenges, including fragmented data and difficulties in data integration.[33] Systematically consolidating existing reports within the field remains a significant challenge. There is a pressing need for databases that encompass high-dimensional data for energy materials. Moreover, it is critical to develop databases that integrate both experimental and theoretical data while supporting interdisciplinary applications. With the recent surge in AI technologies, their applications in scientific research have deepened. However, many existing reported databases lack the incorporation of AI-driven agents, which presents significant challenges for human analysts tasked with processing vast amounts of data. Furthermore, because current datasets often aggregate measurements from different studies, they inevitably suffer from inconsistencies in experimental protocols, variations in reporting standards, and incomplete metadata. Such heterogeneity introduces noise into model training pipelines and substantially limits the reproducibility of AI-driven predictions.[34]

To address these challenges, ongoing efforts are being made to develop more robust and comprehensive databases that integrate both experimental and computational data, ensuring consistency and standardization across materials systems. Advances in AI-driven agents are playing a pivotal role in overcoming data fragmentation and inconsistency by automating data extraction, validation, and curation from diverse sources. These agents can standardize experimental protocols, fill gaps in incomplete metadata, and align data from different studies, thus improving the reliability and usability of databases. Additionally, the closed-loop feedback systems ensure continuous refinement of both data and models. These integrated efforts collectively enhance the reproducibility and reliability of AI-driven predictions, advancing the efficiency of materials discovery in energy and catalysis research.

3. Physical models as the scientific core of modern digital materials

In this section, we focus on the role of physical models in modern digital materials research. These models serve as the scientific foundation for understanding the behavior of materials at different scales, bridging the gap between theoretical predictions and experimental results. By examining key physical models, such as those used in catalysis and battery research, we highlight how they complement the data-driven approaches within the digital materials ecosystem, providing essential insights that guide the design and discovery of new materials.

3.1. Representative physical models in catalysis

Among the wide variety of physics-based approaches developed for electrocatalysis, several key physical models have significantly shaped both our fundamental understanding and practical predictions. These models primarily include surface Pourbaix phase diagrams based on surface state calculations, the common catalytic model (such as free energy diagrams, transition states, and scaling relationships), and pH-dependent volcano models. While not exhaustive, these examples effectively illustrate how theory can bridge atomistic processes with macroscopic catalytic performance.

The first example is the surface Pourbaix diagram, initially proposed by Hansen *et al.*[35] in 2008, as a DFT-based phase

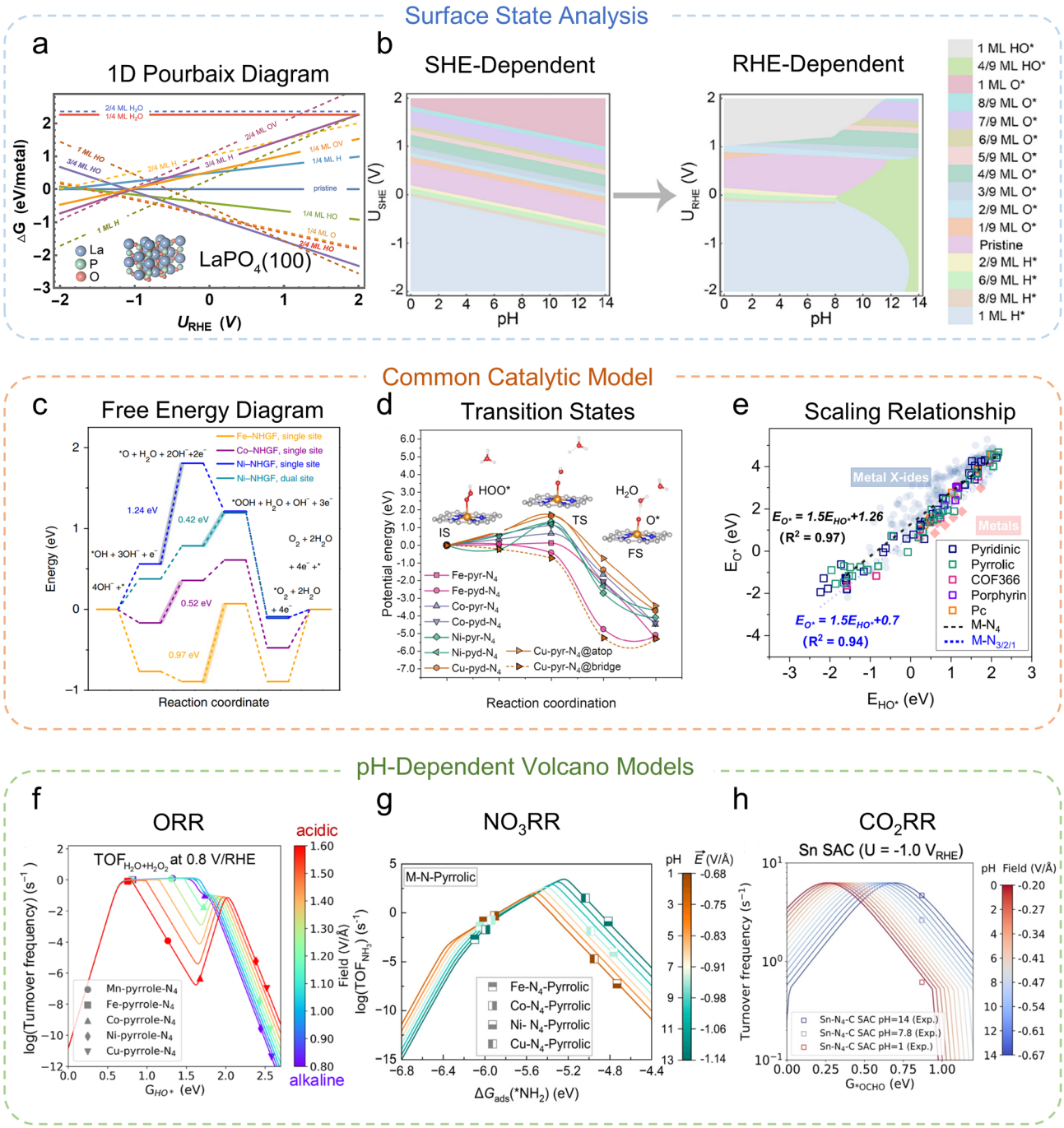

Fig. 3 Representative physical models for catalytic materials. (a) 1D Surface Pourbaix diagram.[36] (b) Classical surface Pourbaix diagram at the standard hydrogen electrode (SHE) scale (left) and the advanced pH-dependent surface Pourbaix diagram at the reversible hydrogen electrode (RHE) scale (right).[37] (c) An example of the application of the energy diagram in the oxygen evolution reaction (OER) to describe reaction thermodynamics.[38] (d) Transition energy barrier calculated from CI-NEB.[46] (e) Scaling relationships for catalytic activity modelling[45] and (f) pH-dependent microkinetic modeling to derive a pH-dependent volcano model for ORR,[45] which was extended to (g) NO_3RR[47] and (h) CO_2RR.[48] Adapted with permission from: (a) ref. 36 © 2023 AIP Publishing, (b) ref. 37 © 2024 The Authors, (c) ref. 38 © 2018 Springer Nature (d) ref. 46 © 2025 The Authors, (e and f) ref. 45 © 2024 The Authors, (g) ref. 47 © 2025 The Authors and (h) ref. 48 © 2025 The Authors.

diagram framework to describe the stability of surface states (*i.e.*, the surface coverage under electrochemical operating conditions) as a function of applied potential and pH. This pioneering idea has since been extended by Liu *et al.*[36] to systematically survey transition metal oxides, carbides, nitrides, and hydroxides (Fig. 3a), demonstrating that the electrochemical *operando* surface states are often drastically different from the pristine stoichiometric structure, thereby highlighting the necessity of preliminary electrochemical surface state verification. More recently, Liu *et al.*[37] advanced this model into

a reversible hydrogen electrode (RHE)-dependent formulation that incorporates electric field corrections and potentials of zero-charge, enabling the accurate prediction of pH-dependent surface coverage (Fig. 3b) and providing a closer bridge between theoretical predictions and experimental observations.

Another widely used example is the electrochemical free energy diagram (Fig. 3c),[38] developed following Nørskov's seminal computational hydrogen electrode (CHE) model.[39] Its simplicity and intuitive mapping of free-energy changes along elementary steps have made it the most applied framework in electrocatalysis. In practice, it is often coupled with kinetic tools such as the nudged elastic band (NEB) method (Fig. 3d), developed by Henkelman *et al.*,[40,41] to capture the activation barriers from the complex potential energy surfaces of atomistic systems. Additionally, scaling relationships (Fig. 3e) are frequently used in the construction of catalytic activity volcano plots to describe the relationships between the adsorption energies of different intermediate species. However, this framework remains fundamentally limited: it is thermodynamics-centered and computationally expensive to fully capture kinetics-dominated electrocatalytic behavior, and lacks explicit treatment of pH effects under the realistic RHE conditions.

To address these challenges, the third example is pH-dependent microkinetic modeling, which explicitly integrates thermodynamics, kinetics, and the electrochemical environment. This approach was pioneered by Kelly and co-workers,[42] who incorporated electric field effects and potential of zero-charge simulations into the CHE framework to rationalize the pH dependence of the ORR on Pt and Au electrodes (Fig. 3f). Li, Nørskov, and colleagues subsequently demonstrated how this methodology explains the intrinsic limitations of transition metal oxides for oxygen reduction reaction (ORR) in hydrogen fuel cell applications[43] and the pH dependence of SACs for electrocatalysis,[44–46] and further extended it to the highly complex nitrate reduction reaction (NO_3RR,[47] Fig. 3g) and CO_2RR[48] (Fig. 3h), showcasing its general applicability. By coupling potential, pH, and coverage effects into kinetic simulations, this modeling method provides a more realistic description of catalytic activity and selectivity under *operando* conditions.

Many modeling approaches in catalysis have the potential to make significant contributions to the digital materials ecosystem. Taken together, the three examples discussed above illustrate how representative physical models have evolved, from static surface thermodynamics to dynamic, pH-dependent kinetic frameworks, laying the foundation for next-generation modeling strategies that aim to predict catalytic behavior with both chemical accuracy and broad applicability.

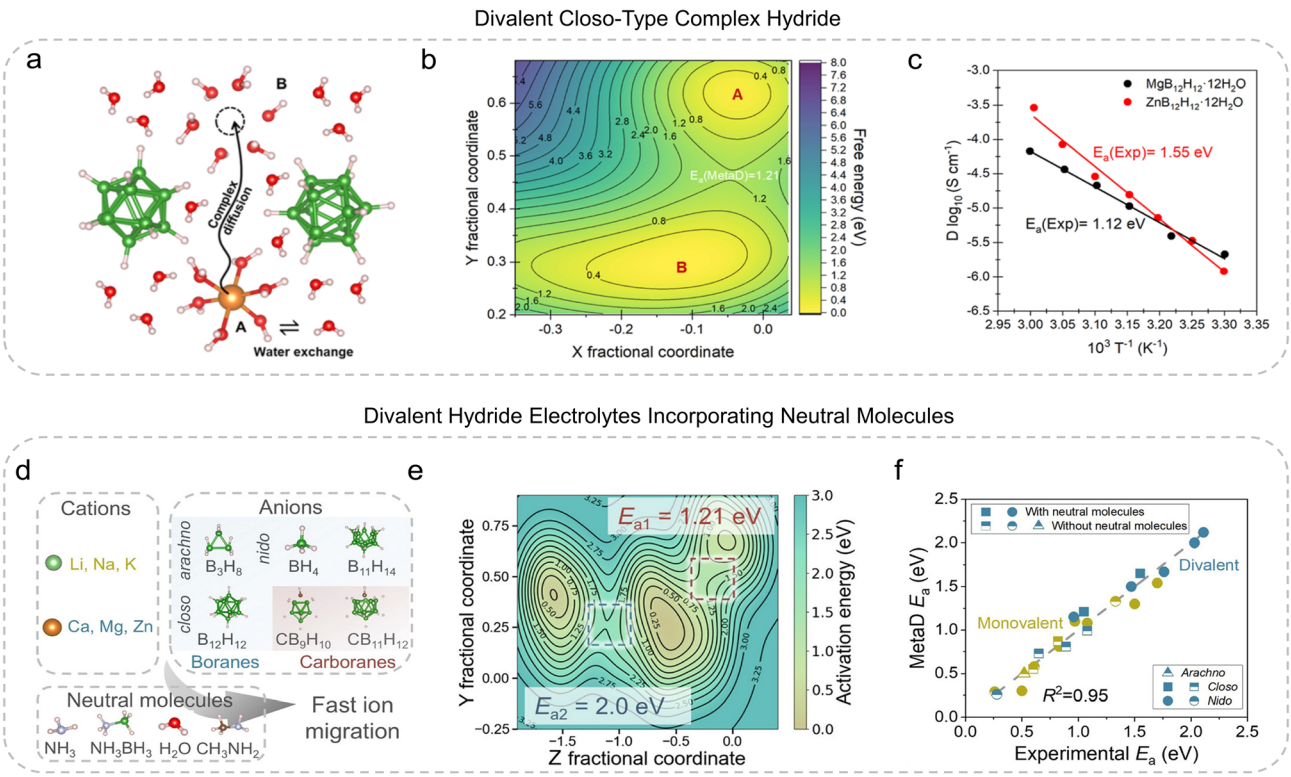

Fig. 4 Representative physical models for SSEs. (a) Migration pathway (direction: A → B) of the $[Mg(H_2O)_x]^{2+}$ hydrocomplex to the next vacant site.[50] (b) Potential energy surfaces for the A → B migration in $MgB_{12}H_{12}\cdot12H_2O$.[50] (c) Experimental conductivity as a function of temperature.[50] (d) Typical cations, anions, and neutral molecules in hydride SSEs.[51] (e) The potential energy surface of $Mg(BH_4)_2\cdot2NH_3$ as captured by MetaD simulations.[51] (f) Comparison of experimental activation energy (E_a) with simulated E_a from MetaD simulations for structures with (filled icons) and without (half-filled icons) neutral molecules.[51] Adapted with permission from: (a–c) ref. 50 © 2023 American Chemical Society and (d–f) ref. 50 © 2025 John Wiley and Sons.

3.2. Representative physical models in batteries

The development of solid-state electrolytes (SSEs) is at the forefront of research aimed at improving the performance and safety of next-generation energy storage systems.[49] Among the materials under investigation, divalent complex hydride-based SSEs, particularly divalent *closo*-type complex hydrides (CTCHs), have garnered attention due to their promising ionic conductivity and high electrochemical stability. However, these materials present unique challenges, such as the strong electrostatic interactions between divalent cations and their counter anions, which hinder ionic migration and impede the design of high-performance electrolytes. To address these challenges, researchers have increasingly turned to advanced computational models that can simulate and predict the behavior of these complex materials, accelerating the development of novel SSE candidates.

Campos dos Santos *et al.*[50] presented a study that integrates a genetic algorithm (GA)-based global optimization method with *ab initio* metadynamics (MetaD) simulations to explore the structure–performance relationships of divalent CTCHs (Fig. 4a). This integrated computational strategy allows for the prediction of stable crystal structures and cation diffusion activation energies (E_a) without relying on experimental data (Fig. 4b). By combining GA and MetaD, the study successfully predicted structural information and activation energies that were in excellent agreement with experimental observations (Fig. 4c). This approach not only unveiled the impact of neutral molecules, such as water, on the ionic conductivity of CTCHs but also identified key factors that promote cation diffusion, ultimately providing insights into the design of more efficient SSEs for battery applications.

A notable study by Wang *et al.*[51] introduced an innovative data-driven AI framework that integrates LLMs, *ab initio* MetaD simulations, and multiple linear regression to explore and predict the migration mechanisms of hydride SSEs (Fig. 4d). The research highlights a novel "two-step" migration model that involves an initial "coordination-unlock" stage, followed by a "paddle-wheel" mechanism. This process was observed in the migration of Mg^{2+} ions in $Mg(BH_4)_2 \cdot 2NH_3$ and Li^+ ions in $LiBH_4 \cdot NH_3$ (Fig. 4d). These findings are significant as they demonstrate how neutral molecules, such as NH_3, facilitate ionic migration by disrupting the strong electrostatic interactions that typically hinder divalent ion movement. The presence of these neutral molecules within the SSE lattice results in a substantial decrease in activation energy (E_a), as demonstrated by the close agreement between the experimental E_a and the MetaD-simulated E_a values (Fig. 4f).

The application of advanced physical models integrated with GA and MetaD simulations has proven to be an invaluable approach in the development and optimization of SSEs. These models offer deep insights into the structure–performance relationships of complex materials, particularly those involving divalent ions, by accurately predicting cation diffusion mechanisms and activation energies. The continued evolution of the digital materials ecosystem, powered by key simulation methods, will be critical in the development of advanced solid-state batteries.

4. Machine intelligence for materials discovery

In this section, we explore the integration of machine intelligence into the materials discovery process. As part of the digital materials ecosystem, machine learning and AI algorithms are transforming the way materials properties are predicted, discovered, and optimized. By examining how these techniques are applied to systems such as thermoelectrics, electrocatalysts, and hydrogen storage materials, we demonstrate how machine intelligence accelerates the identification of promising new materials and enhances the efficiency of the discovery process.

4.1. ML for the direct prediction of thermoelectric properties

Thermoelectric (TE) conversion technology is a promising approach for waste heat recovery and solid-state refrigeration, as it enables direct heat-to-electricity conversion through the Seebeck and Peltier effects.[52,53] The dimensionless figure of merit (zT) quantifies the performance of a TE material, where σ, S, κ_e, κ_L, and T denote the electrical conductivity, Seebeck coefficient, electronic thermal conductivity, lattice thermal conductivity, and absolute temperature, respectively. The term σS^2 represents the power factor. In recent years, because TE materials have rather straightforward structure–property relationships, the rapid development of ML has facilitated the exploration of novel TE materials, leading to substantial progress in this field. Therefore, we use TE materials as the key example for ML regression in this section. Supervised learning has been extensively employed to establish relationships between features (*e.g.*, compositional descriptors) and target values, such as TE properties, for both regression and classification tasks. Regression models,[54] in particular, have been developed to directly predict properties including the S,[55–58] κ_L,[59–62] σS^2,[63–65] and zT[66–68] across various materials systems, such as 2-1-2 Zintl phases, half-Heusler compounds, and broader TE systems. In several cases, these predictions have been experimentally validated. For example, Er_2Te_3 (ref. 62) exhibits an ultralow κ_L of ~0.5 W m^{-1} K^{-1} at 937 K, which was first determined by integrating a crystal graph convolutional neural network (CGCNN) with random forest models to predict κ_L. Upon carrier concentration optimization, $Er_2Te_{2.7}Bi_{0.3}$ achieved a zT of ~1.0 at 973 K. More importantly, in developing the regression model for predicting temperature-dependent target values, a composition-based cross-validation strategy was introduced[67,69] (Fig. 5a). Critically, this approach emphasizes that data points with the same composition but different temperatures should not be split into separate sets to avoid overfitting. By applying this data splitting method, the ML model achieved an excellent R^2 value and enabled reliable predictions of zT for unexplored materials. Subsequent DFT calculations revealed maximum zT values of 1.98 and 2.12 for n- and p-type $Ge_2Te_5As_2$, and 0.58 and 0.74 for n- and p-type $Ge_3(Te_3As)_2$, respectively, highlighting their promise as thermoelectric materials. In addition to regression, classification models have been employed in the exploration of potential TE materials. Sun *et al.*[68] employed deep neural networks to

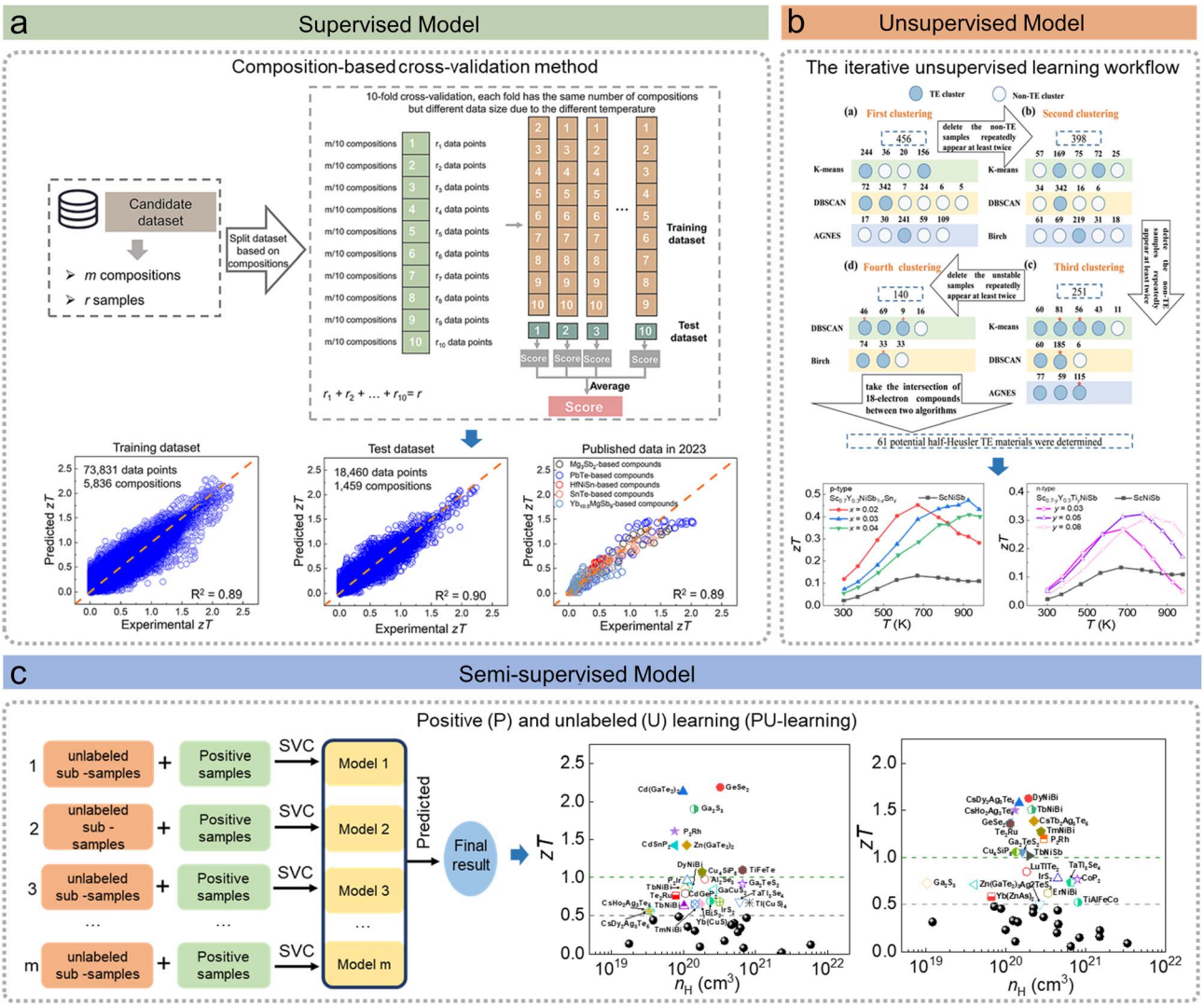

Fig. 5 Examples of machine learning (ML) models for materials property predictions, including supervised models, unsupervised models, and semi-supervised models. (a) Schematic illustration of the composition-based 10-fold cross-validation strategy, and the comparison of ML model performance across the training set, testing set, and an independent dataset published in 2023.[67] (b) Iterative integration of unsupervised ML with labeled and reported half-Heusler TE materials, and the ScNiSb-based TE material was experimentally investigated;[71] (c) PU learning framework employing bootstrap aggregating (bagging) techniques, and the identified potential TE materials were validated through theoretical calculations.[73] Adapted with permission from: (a) ref. 67 © 2024 Springer Nature, (b) ref. 71 © 2022 The Authors and (c) ref. 73 © 2023 AIP publishing.

construct both a regression model for predicting the maximum zT value and a classification model to determine the corresponding optimal doping type. Through this approach, $Pb_2Sb_2VI_5$ (VI═S, Se, Te) compounds were identified and validated by theoretical calculations. Moreover, classification models can also be utilized to categorize TE materials into binary classes, such as high *versus* low S or σ, by appropriately adjusting and confirming threshold values,[70] resulting in a list of possible TE material candidates.

Unsupervised learning does not require well-labeled training data, whereas semi-supervised learning relies on partially labeled datasets. Both approaches have garnered attention in situations where supervised learning is challenging due to the scarcity of sufficient labeled data. For example, unsupervised clustering methods,[71] including K-means and Gaussian Mixture Models, were employed to group half-Heusler compounds into distinct clusters based on generated features (Fig. 5b). ScNiSb was identified as a promising candidate, and subsequent experiments achieved peak zT values of ~0.5 at 925 K in p-type $Sc_{0.7}Y_{0.3}NiSb_{0.97}Sn_{0.03}$ and ~0.3 at 778 K in n-type $Sc_{0.65}Y_{0.3}Ti_{0.05}NiSb$. Unsupervised word embeddings[72] trained on materials literature can capture latent knowledge and be applied to explore potential TE materials. Predictions derived from historical literature have been validated by recently reported TE materials, demonstrating that such insights can effectively guide the discovery of new candidates. Positive and unlabeled (PU) learning,[73] a semi-supervised learning method, was proposed to train a classifier to distinguish reported TE materials (P) from unreported materials (U) (Fig. 5c). Using this approach, the probabilities of unlabeled materials belonging to the TE class were predicted. Finally, forty candidate TE materials were identified. Eight p-type and twelve n-type materials

exhibited excellent theoretical zT values greater than 1. In addition, a semi-supervised generative learning framework was developed for the inverse design of TE materials,[74] combining limited labeled data with augmented unlabeled data to generate and validate high-performance candidates. The designed compound $Mg_{3.1}Sb_{0.5}Bi_{1.497}Te_{0.003}$ exhibited a zT of 0.75 at 300 K, surpassing most known inorganic materials at room temperature. Therefore, ML has become an indispensable tool for TE materials research, enabling efficient property prediction, data-driven screening, and inverse design.[75–77]

4.2. ML for catalyst performance predictions

In the field of electrochemical energy technologies, electrocatalysis plays a pivotal role in accelerating reaction kinetics through the use of catalyst materials.[78] With the continuous advancement of experimental techniques and theoretical simulations, a vast amount of electrocatalyst-related data has been accumulated, providing a solid foundation for the application of ML in the exploration and design of high-performance electrocatalysts. Based on experimental data, ML has been employed to predict the properties of various electrocatalysts under different electrochemical reactions, such as overpotentials for the OER in V-doped Ni–Co layered double hydroxides[79] and NiCoFe oxide catalysts,[80] as well as current densities for the ORR in multicomponent metal oxides (Fig. 6a).[81]

From a theoretical perspective, the interaction between reaction intermediates and catalyst surfaces plays a crucial role in determining catalytic performance. The adsorption energies of key intermediates are typically calculated to quantify their

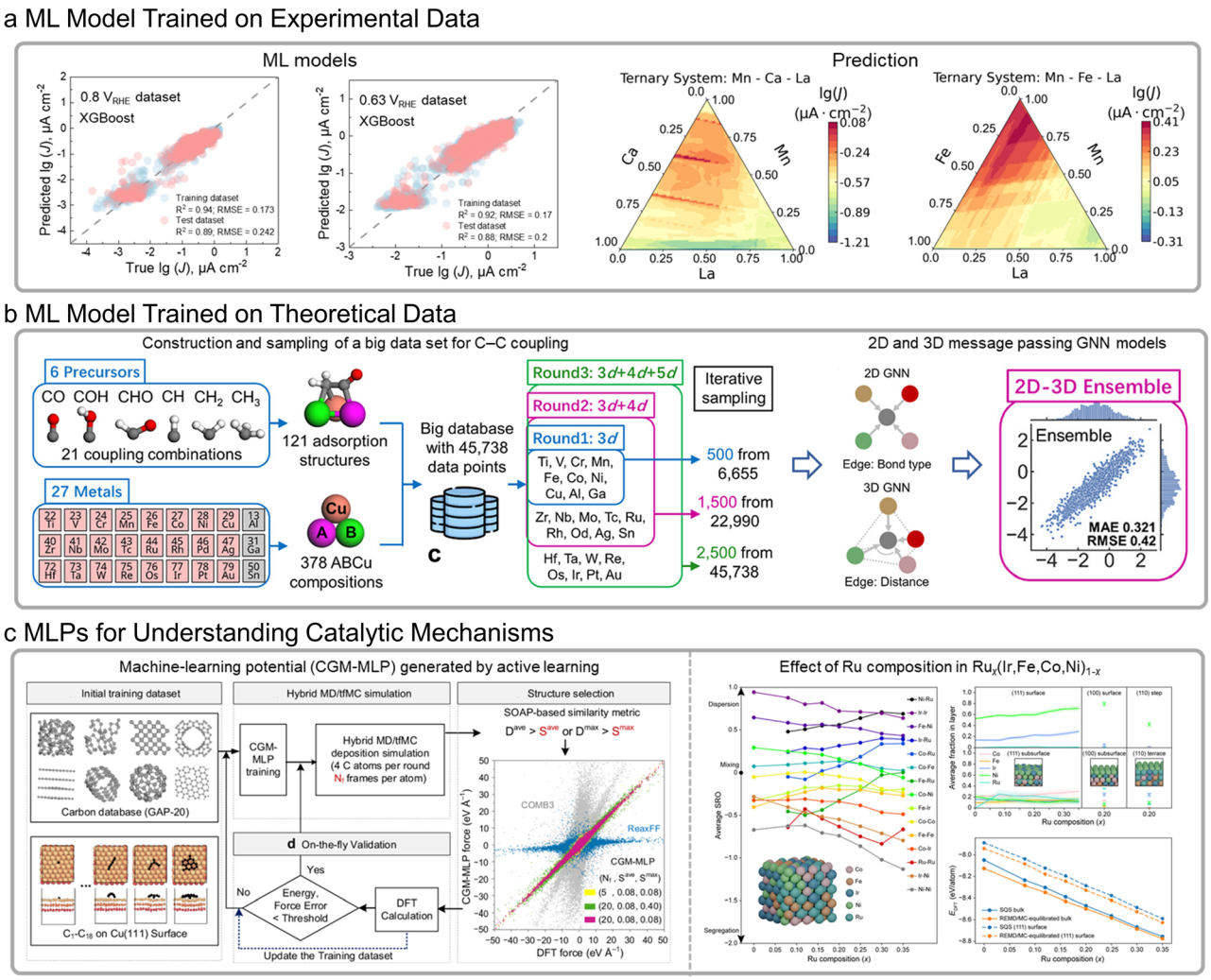

Fig. 6 Examples of machine learning (ML) for catalyst performance prediction, including models trained on experimental and theoretical data, and those employing ML potentials (MLPs). (a) Diagonal scatter plot comparing experimental and predicted values by XGBoost on the training and test sets at 0.8 and 0.63 V_{RHE}, along with a contour map illustrating the model-predicted current densities across different multicomponent metal oxides.[81] (b) Development of the 2D–3D ensemble model for C–C coupling big data set prediction.[84] (c) Left: schematic illustration of MLP generated by active learning on-the-fly during hybrid molecular dynamics and time-stamped force-biased Monte Carlo (MD/tfMC) simulations.[89] Right: theoretical analyses combining a MLP with replica exchange molecular dynamics and Monte Carlo based atom swaps (REMD/MC) for understanding catalytic behavior.[90] Adapted with permission from: (a) ref. 81 © 2024 The Authors, (b) ref. 84 © 2024 American Chemical Society, and (c) ref. 89 © 2025 The Authors and ref. 90 © 2025 American Chemical Society.

binding strengths with the surface. These energetic parameters serve as fundamental descriptors for constructing microkinetic volcano models that predict theoretical catalytic activities. Therefore, obtaining accurate adsorption energies is essential for assessing electrocatalytic performance. ML methods have been employed to predict adsorption energies for key intermediates across a vast number of catalytic sites. For instance, ML models have been used to evaluate the adsorption energies of OH* on millions of reactive sites of different crystal facets in high-entropy alloys (HEAs) for ORR,[82] to predict CO adsorption energies on layered alloy surfaces relevant to CO_2-to-methanol conversion,[83] and to estimate the adsorption energies of six C_1 precursors (CO, COH, CHO, CH, CH_2, and CH_3) and twenty-one C_2 combinations (six symmetric and fifteen asymmetric couplings) involved in the C–C coupling processes (Fig. 6b).[84,85]

In addition, MLPs are typically trained on datasets generated from DFT calculations, where total energies and atomic forces serve as the training targets.[86,87] By learning the relationships between local atomic environments and these physical quantities, MLPs can accurately construct potential energy surfaces, thereby accelerating adsorption energy calculations and providing valuable insights into catalytic mechanisms. For example, the AdsorbML framework[88] and an active-learning Gaussian Approximation Potential (GAP) model (Fig. 6c, left)[89] have been proposed to accelerate adsorption energy evaluations and efficiently identify global minima with formation energies. Furthermore, an MLP coupled with replica-exchange molecular dynamics was employed to describe the effect of Ru composition variation on phase formation and stability in $Ru_x(Ir, Fe, Co, Ni)_{1-x}$ multicomponent alloys under acidic OER conditions (Fig. 6c, right).[90] Collectively, these efforts have greatly advanced the data-driven design and discovery of promising electrocatalysts.[91]

4.3. ML for hydrogen storage property prediction

Recent advances in data-centric materials informatics are transforming the way we design hydrogen storage materials,[92]

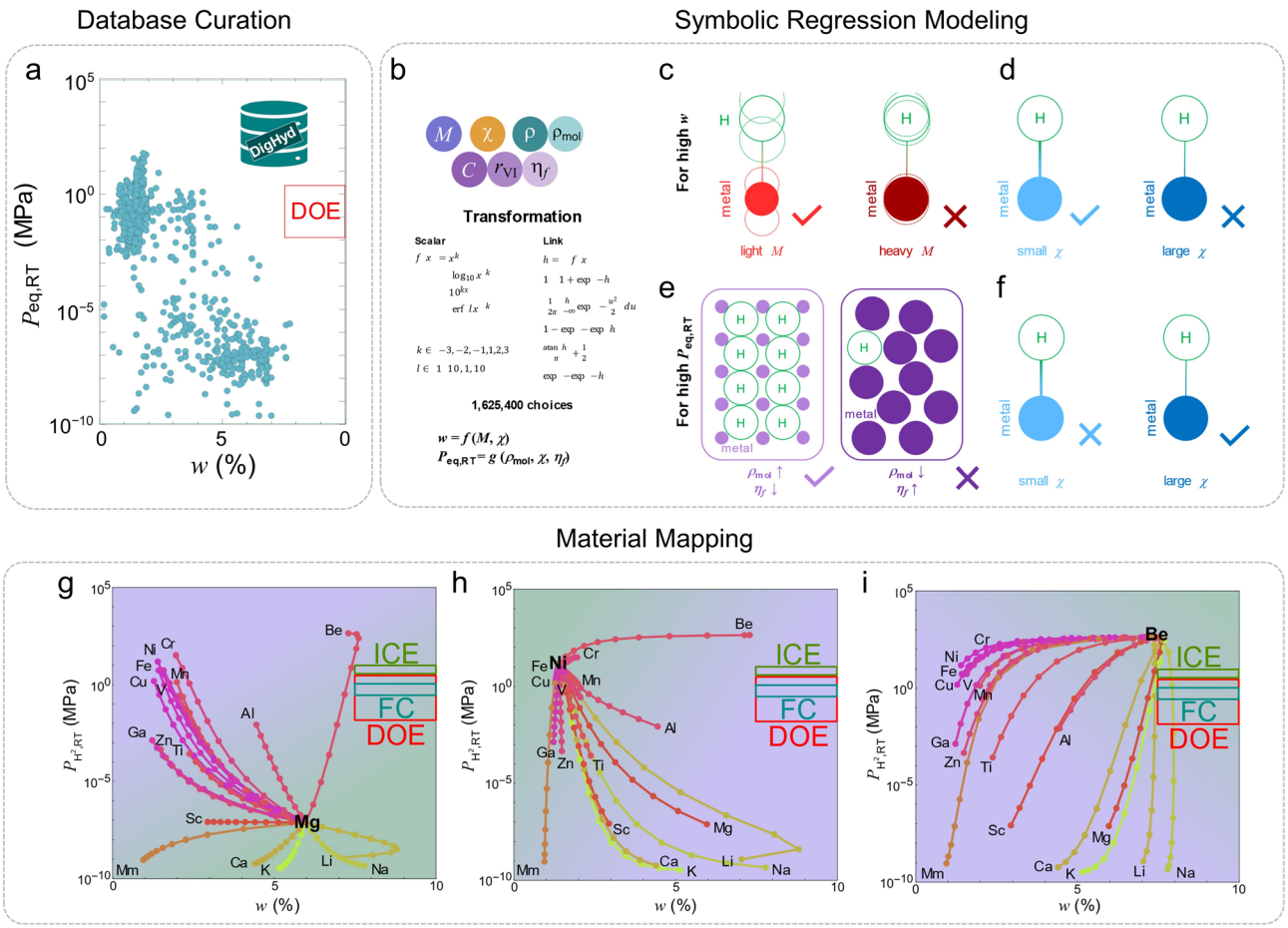

Fig. 7 Integrated simulation perspective for physically interpretable ML modeling of hydrogen storage properties. (a) Overview of the *DigHyd* database showing the broad distribution of equilibrium pressure $P_{eq,RT}$ and gravimetric capacity w across reported metal hydrides, illustrating the unavoidable trade-off between the two properties and the gap from US-DOE targets. (b) Framework of symbolic-regression modeling, where combinations of chemically meaningful descriptors and nonlinear transformations were systematically searched to construct millions of candidate equations. (c–f) Schematic interpretation of key descriptors governing w and $P_{eq,RT}$. (g–i) Descriptor-based design maps generated from the regression models for compositions anchored on Mg, Ni, and Be, respectively. The maps visualize compositional pathways linking saline- and interstitial-type hydrides and highlight that Be-containing systems, especially Be–Na alloys, uniquely approach the US-DOE target zone (red = ultimate, green = internal combustion engine, and blue = fuel cell). Reproduced from ref. 94, under the terms of the Creative Commons CC BY-NC license.

particularly metal hydrides.[93] Building on the Digital Hydrogen Platform (*DigHyd*),[31] which consolidates over 30 000 experimentally curated pressure–composition isotherms, Jang *et al.* demonstrated how physically interpretable ML can be harnessed to predict key thermodynamic and gravimetric metrics, gravimetric hydrogen density (w) and the equilibrium pressure at room temperature ($P_{eq,RT}$), with accuracy comparable to that of state-of-the-art black-box models. Fig. 7 presents an integrated view of their simulation-based framework that connects large-scale data curation, symbolic-regression modeling, and descriptor-guided materials mapping. Fig. 7a summarizes the *DigHyd* database, which aggregates thousands of experimentally measured pressure–composition isotherms for metal hydrides. The scatter of equilibrium pressure $P_{eq,RT}$ *versus* gravimetric capacity w demonstrates an inherent performance trade-off: light-element hydrides achieve high w but exhibit excessively low pressures, whereas transition-metal hydrides release hydrogen readily but with low capacity. None of the known compositions reach the US-DOE Target window, emphasizing the need for predictive models that can transcend existing empirical limits. Fig. 7b outlines the symbolic-regression approach used to derive physically interpretable equations for $w$and $P_{eq,RT}$. Starting from four chemically intuitive descriptors (atomic mass M, electronegativity χ, molar density ρ_{mol}, ionic filling factor η_f, *etc*),[94] they performed a high-throughput search over more than a million candidate analytical forms generated by combining scalar and link transformations. This exhaustive exploration yielded compact closed-form models that match the accuracy of black-box ML regressors while preserving transparent physical meaning. The outcome demonstrates that even complex hydrogen-storage behavior can be captured by simple, human-readable equations when descriptor selection and model search are systematically orchestrated. Fig. 7c–f schematically illustrates how each descriptor affects storage performance. For high capacity, low atomic mass and strong bond polarity (large χ) are beneficial because they reduce lattice weight and strengthen metal atom–hydrogen interactions. In contrast, high equilibrium pressure, desirable for room-temperature hydrogen release, is favored by densely packed lattices with low ionic filling factors and weaker bond polarity (small χ). The fact that χ exerts opposite influences on w and $P_{eq,RT}$ explains the persistent trade-off observed in panel (a) and frames the chemical origin of the capacity–pressure dilemma. Finally, Fig. 7g–i shows descriptor-based design maps derived from the regression equations. The Mg-anchored pathway typifies saline-type hydrides, offering high w but low $P_{eq,RT}$; the Ni-anchored pathway represents interstitial hydrides with the reverse trend. Remarkably, the Be-anchored map reveals a distinct trajectory that approaches the US-DOE target region,[95] identifying Be and its alloys (particularly Be–Na) as unique compositions capable of balancing both metrics. This "bird's-eye" view underscores the predictive and explanatory power of the symbolic-regression framework, which transforms large experimental datasets into quantitative, physically interpretable guidance for the rational design of next-generation hydrogen-storage materials.

5. LLM-based AI agents for modern digital materials: from knowledge extraction to design

Developed based on LLMs, AI agents are playing an increasingly important role in materials development,[96] primarily in two areas: first, by accelerating the extraction and structuring of materials data to build comprehensive databases, and second, by formulating scientific hypotheses and research plans to guide material discovery. These agents also facilitate the design of new materials based on those databases and predictive models.[97] The following sections present three typical examples of AI agents applied to hydrogen storage materials and battery materials, highlighting how these tools are transforming the materials research landscape.

5.1. AI agents for database development

In database construction, the recently developed descriptive interpretation of visual expression (DIVE)[31] workflow introduces a detailed process for extracting key material properties that are often presented only in literature figures. For hydrogen storage materials, these typically include PCT curves, TPD curves, and discharge curves. First, a lightweight inference model scans the figure captions of a paper to determine whether these key figures are present. When such figures are detected, the corresponding figure, its caption, and the relevant surrounding text are passed to a second multimodal LLM. Through carefully designed prompts, the LLM is guided to extract the key points from each curve, positioning them correctly in the output (Fig. 8b, Prompt Design). The extracted textual data then replace the original figure in the article. Finally, the modified article—now with figures represented as text—is passed to a third LLM for the final extraction of numerical and contextual data. Ansari and Moosavi's Eunomia framework[98] is an LLM-based literature agent that autonomously converts full-text articles into structured, ML-ready materials datasets. It reads papers, identifies entities such as hosts, dopants and compositions, and extracts quantitative relationships, achieving near-state-of-the-art performance on several extraction tasks without task-specific training. Odobesku *et al.*[99] generalized this idea to multimodal information extraction by coordinating multiple agents that jointly process text, figures, and captions. Their system combines vision models, OCR and language models to interpret plots and micrographs, link them to the surrounding prose, and fuse everything into coherent records of synthesis conditions, structures, and properties for materials science.

5.2. LLMs for design insight in materials discovery

Data-driven frameworks combined with physical modeling have emerged as powerful tools to accelerate SSE discovery. Among these, the DDSE database for SSEs represents a milestone, establishing a unified and evolving platform that integrates experimental data, computational modeling, and comprehensive AI tools for the systematic understanding and prediction of SSE performance.

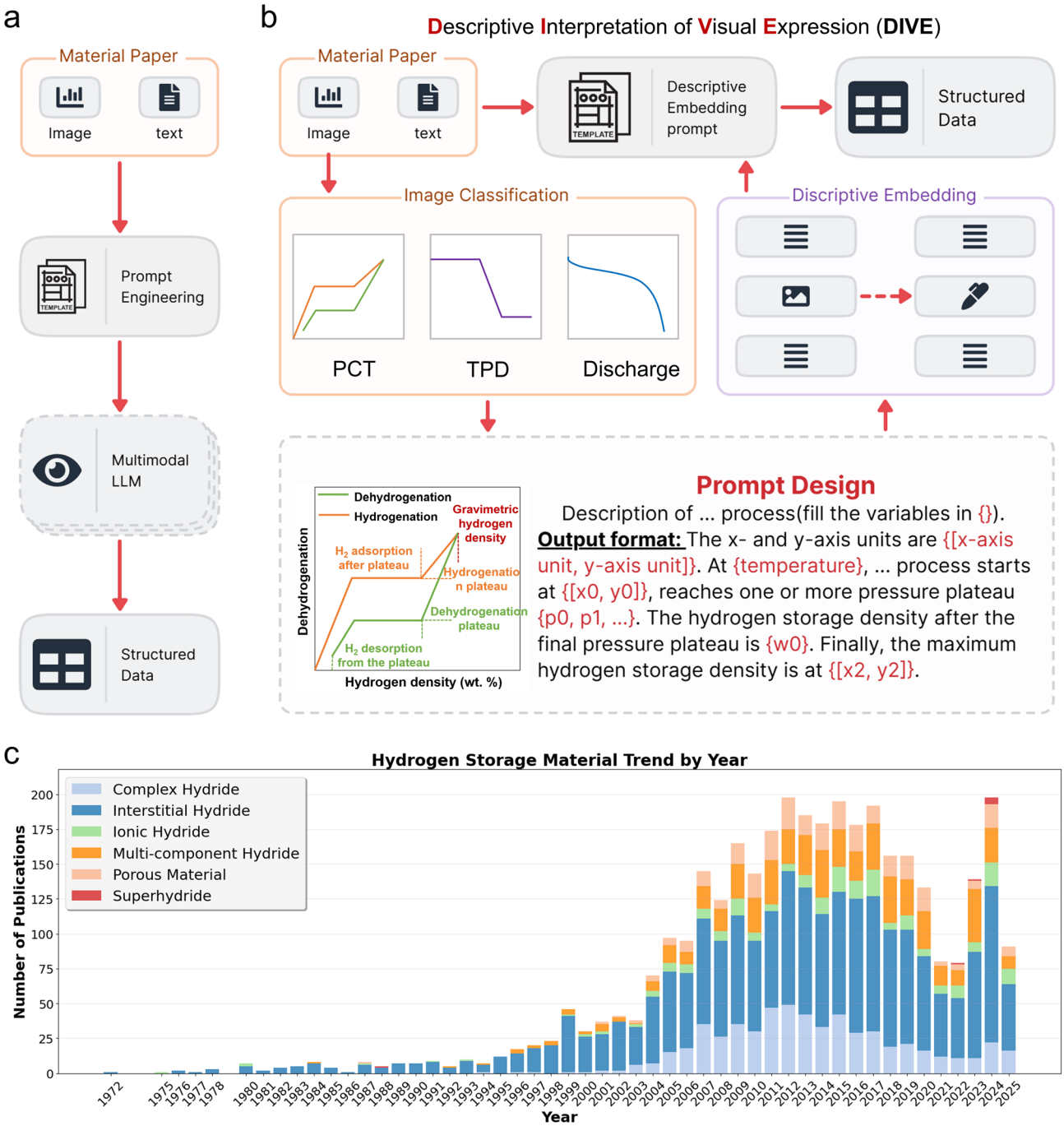

Fig. 8 Multimodal data extraction pipeline: the descriptive interpretation of visual expression (DIVE).[31] (a) Conventional extraction pipeline based on a single multimodal LLM. (b) DIVE extraction pipeline, where descriptive prompts embed key data points and generate image replacements for structured data extraction. (c) Annual publication trends categorized by different types of hydrogen storage materials. Reproduced from ref. 31, under the terms of the Creative Commons CC BY-NC license.

The DDSE (now renamed as *DigBat*: https://www.digbat.org), developed by Li and co-workers, compiles a comprehensive and large-scale dataset of SSEs.[100] As of February 2026, it contains approximately 3000 experimental materials, 25 996 ionic conductivity measurements and 863 computational entries. Moreover, the DDSE functions as a dynamic and self-improving research infrastructure that continuously incorporates new data from both experiments and simulations. Automated data extraction and model standardization enable rapid exploration of ionic conductivity trends across wide chemical and structural spaces. By coupling large-scale statistical analysis with LLMs, DDSE identifies transport-related parameters such as activation energy, temperature dependence, and carrier type, thereby linking physical descriptors with measurable macroscopic properties. Importantly, DDSE supports iterative model

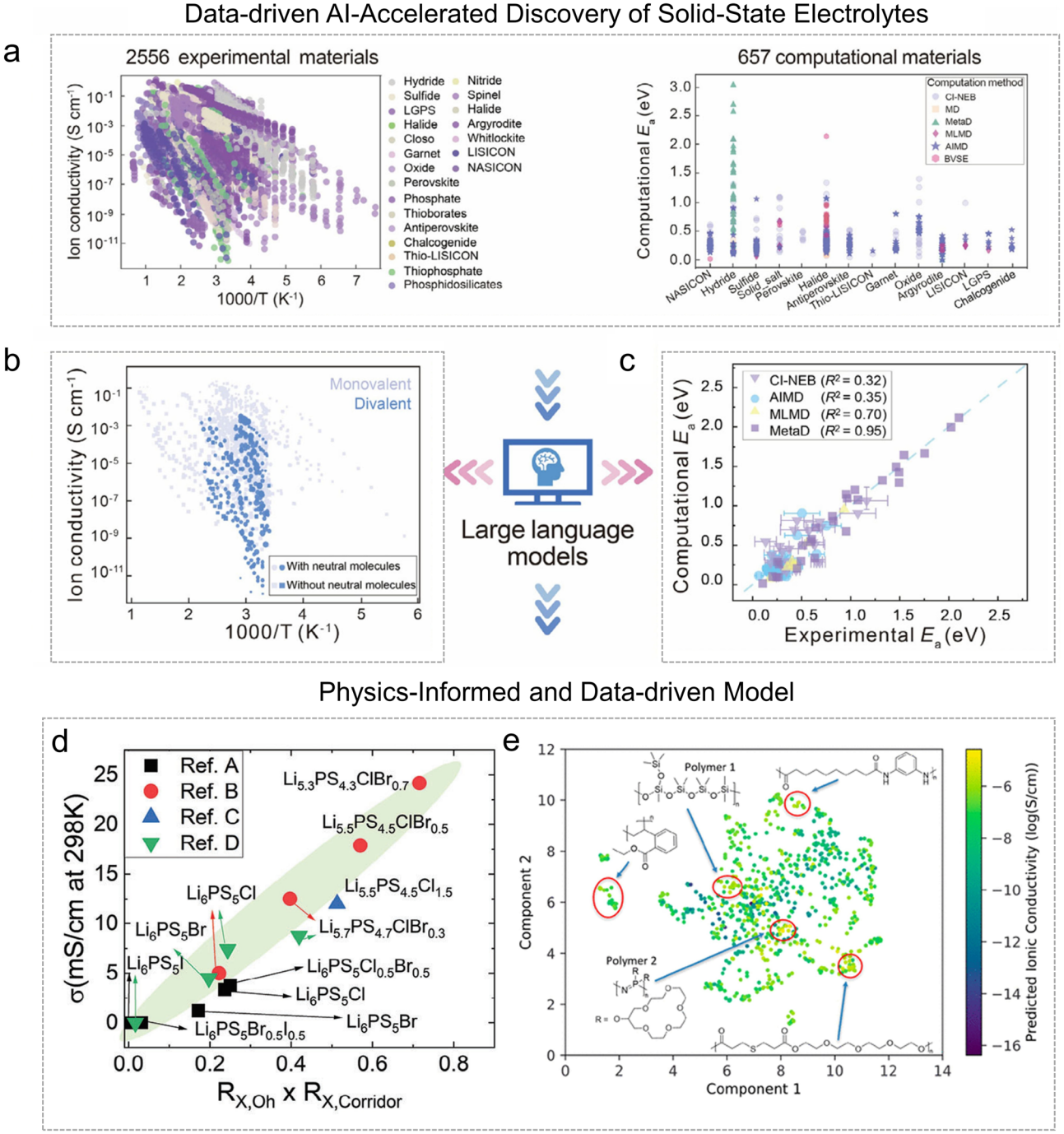

Fig. 9 Data-driven, AI-accelerated discovery of solid-state electrolytes (SSEs). (a) Ion conductivity *vs.* inverse temperature for ~3000 experimental materials and activation energies for ~700 computational materials.[51] (b) Conductivity *vs.* inverse temperature for monovalent and divalent SSEs, with/without neutral molecules.[51] (c) Comparison of computational and experimental activation energies for different methods.[51] (d) Ionic conductivity at 298 K *vs.* structural descriptors for sulfide-based SSEs.[104] (e) Predicted ionic conductivity of polymer electrolytes based on structural components.[105] Adapted with permission from: (a–c) ref. 51 © 2025 John Wiley and Sons, (d) ref. 104 © 2024 John Wiley and Sons and (e) ref. 105 © 2023 The Authors.

refinement through feedback between simulation and experiment, transforming it from a static database into a dynamic predictive engine (Fig. 9a). This workflow enhances the interpretability of data-driven models and enables physics-informed correlation mapping, bridging the gap between first-principles accuracy and experimental observability.

Metal hydride-based SSEs provide a representative demonstration of DDSE's predictive capabilities.[101] Metal hydrides possess light-element frameworks, flexible lattice structures, and tunable cation–anion interactions, making them ideal systems for model-driven discovery. Using the DDSE data, a large-scale analysis of divalent hydrides containing neutral

molecules was conducted by combining big-data analytics with LLM-assisted feature extraction to reveal how lattice coordination and molecular incorporation affect cation migration.[51,101] The results revealed two universal insights. First, the inclusion of neutral molecules such as NH_3 promotes divalent ion migration by reducing electrostatic confinement and increasing the dynamic reorientation of BH_4^- clusters (Fig. 9b). Second, a consistent gap was observed between experimental and simulated activation energies, indicating that traditional static simulations often neglect configurational entropy effects (Fig. 9c).

Beyond the DDSE framework, numerous researchers have independently applied physics-informed and data-driven models to other classes of SSEs, revealing diverse yet convergent mechanisms of ionic transport. For antiperovskites (X_3BA, X = Li, Na), descriptor-based learning identified the ratio between the tolerance factor and atomic packing factor as a negative predictor of ionic conductivity. Using this compact descriptor, nitro-halide double antiperovskites such as Li_6NClBr_2 and Li_6NBrI_2 were predicted to reach room-temperature conductivities above 1×10^{-4} S cm^{-1} in AIMD simulations.[102] In the famous garnet-type oxides ($Li_7La_3Zr_2O_{12}$, LLZO), data-mining combined with molecular dynamics revealed that Ga^{3+} occupation of octahedral sites enhances Li^+ migration, while Sc^{3+} co-doping promotes redistribution between octahedral and tetrahedral sites.[103] However, the non-monotonic conductivity trend in Ga/Sc co-doped LLZO illustrates the complex balance between carrier concentration and mobility, underscoring the importance of mechanistic frameworks beyond empirical fitting. For sulfide-type argyrodites, both experimental measurements and ML modeling demonstrated that ionic conductivity scales linearly with the product of halogen substitution ratios at octahedral and corridor cage centers ($R_{X,Oh} \times R_{X,Corridor}$)[104] (Fig. 9d). Controlled halogen substitution enhances Li^+ mobility by weakening sulfur localization, while excessive doping produces insulating by-products such as LiCl that reduce overall performance. In polymer electrolytes, the ChemArr model integrated the Arrhenius relationship into a predictive neural network, achieving near-experimental accuracy across more than 200 studies and screening over 20 000 polymers.[105] The model identified siloxane- and phosphazene-derived polymers as promising high-conductivity materials with low glass-transition temperatures (Fig. 9f).

Beyond bulk transport behavior, interfacial processes remain a key bottleneck in the development of solid-state batteries. At Li-metal anodes, instability often results from dendrite formation and side reactions. ML models based on support vector machine (SVM) and kernel ridge regression (KRR) identified Sc^{3+} and Ca^{2+} as effective dopants, capable of forming stable SSE interphases that mitigate interfacial degradation.[106] On the cathode side, interfacial resistance is largely governed by the complex microstructure of composite electrodes. Hwang *et al.* introduced advanced microstructural characterization based on semantic segmentation of electron micrographs, enabling automated quantification of porosity, particle distribution, and phase connectivity.[107]

The use of LLMs to analyze data and gain insights into materials is a key approach in advancing digital material ecosystems. By combining data-driven analysis with physics-informed ML models, this method enhances the design and prediction of materials with desired properties. In SSEs, LLMs help identify important relationships between atomic structure and material performance, revealing new trends that guide the development of high-conductivity materials. This process not only refines ML models but also fosters a dynamic, evolving research infrastructure, where continuous feedback between simulations, experiments, and model refinements accelerates material discovery and innovation. This approach ultimately helps build a more connected and efficient materials research ecosystem.

5.3. AI agents for materials design

For materials design, we can apply a similar prompting strategy but explicitly instruct the AI Agent to propose new, unreported compositions. Under these conditions, the *DigHyd* system (https://www.dighyd.org)[31] demonstrated an iterative design–prediction–optimization capability (Fig. 10). In this workflow, researchers can direct the AI agent to generate novel materials by specifying the material class, candidate elements, and target properties such as gravimetric hydrogen density, pressure, and temperature (Fig. 10a).

In the first round, drawing on both its local knowledge base and the analytical, reasoning, and predictive abilities of LLMs, the *DigHyd* Agent proposed $CaMgFe_2$ (Fig. 10b). This candidate was then evaluated using the ML regression model, which predicts hydrogen density directly from the material's composition. With an R^2 value of 0.87, this model provides a reliable first-pass screening for LLM-generated candidates (Fig. 10c). $CaMgFe_2$ was predicted to store 2.64 wt% hydrogen (Fig. 10d).

The AI agent next suggested increasing the Mg content, yielding Mg_2Fe, with a predicted capacity of 4.13 wt%. However, literature reports indicate that Mg_2Fe undergoes hydrogenation and dehydrogenation only at elevated temperatures (300–400 °C), thus failing to meet the design criteria. In response, *DigHyd* refined the composition to $Mg_2Fe_{0.75}Co_{0.25}$, and later to $Mg_2Fe_{0.6}Co_{0.2}Mn_{0.2}$. The latter was predicted to achieve 4.19 wt% hydrogen capacity, with Mn (or alternatively Al) contributing to hydride stabilization and plateau-pressure optimization. Importantly, this final composition has not been reported in any existing database. Together, these results (Fig. 10d) highlight the ability of the *DigHyd* Agent to rapidly design, predict, and iteratively refine material candidates according to user-defined goals—within minutes. If such AI-driven agents are integrated with high-throughput experimental platforms, the efficiency of materials discovery and development could reach an unprecedented level.

In summary, the integration of AI agents into materials research is reshaping the conventional paradigm of discovery.[108] By bridging data extraction, knowledge reasoning, and material design, such agents not only accelerate the pace of research but also enable a deeper understanding of structure–property relationships that were previously difficult to capture. The examples of DIVE and *DigHyd* demonstrate how

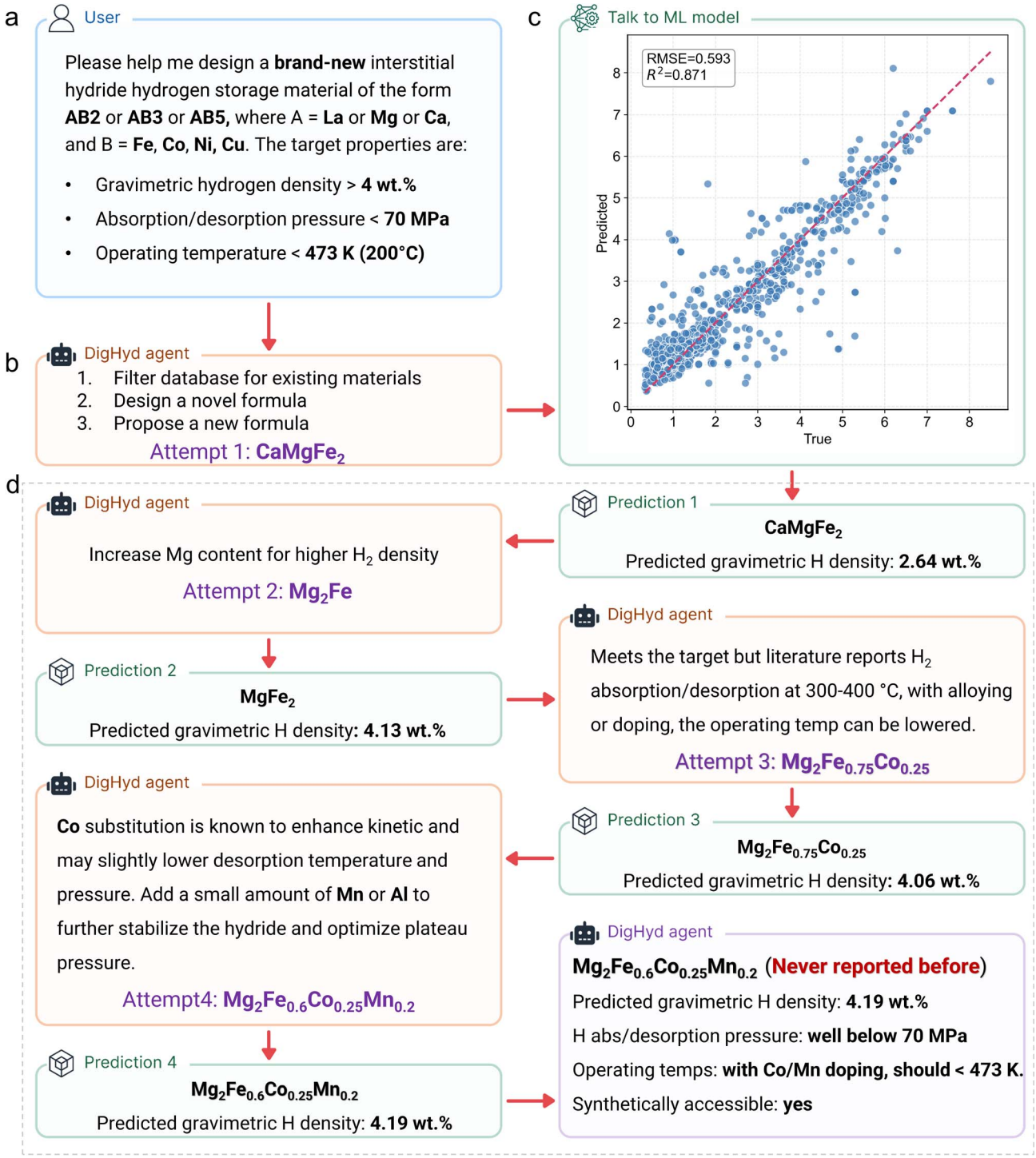

Fig. 10 Workflow of AI agent-driven discovery of new hydrogen storage materials. (a) The user specifies key requirements, including material type, constituent elements, and performance targets. (b) The *DigHyd* agent proposes initial candidate compositions based on data mined from over 4000 historical publications. (c) The candidate compositions are evaluated using a pretrained ML model to predict their gravimetric hydrogen density. (d) *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. Reproduced from ref. 31, under the terms of the Creative Commons CC BY-NC license.

multimodal and generative AI can work "hand in hand"—transforming unstructured literature into structured knowledge and transforming that knowledge into actionable design hypotheses. Looking ahead, the close coupling of AI agents with autonomous experimental platforms will pave the way for a truly self-driving laboratory,[109,110] where materials discovery evolves from a manual and time-consuming process into an intelligent, iterative, and self-improving cycle.

6. Closed-loop: modern digital materials toward autonomous laboratories

Automation of experiments and intelligent selection of materials carry the potential to significantly accelerate scientific discoveries by orders of magnitude.[111–113] Recent studies have attempted to use AI-tools towards smart automation.[5,112,114–116] However, due to small datasets, the use of a single set of algorithms suffers from narrow scope and limited scientific depth.[114,115] To address these challenges, digital materials platforms' (*e.g.*, *DigMat*'s) unique AI-driven, LLM-integrated automated design workflow incorporates vast multidisciplinary experimental datasets and a comprehensive set of physically grounded tools to guide both automated experiments and the user towards novel catalyst design (Fig. 11). Furthermore, the inclusion of cloud-based tools promotes reproducibility and fosters collaboration worldwide (reference place holder for cloud synthesis). User generated experimental data fed back to the system closes the loop resulting in a self-improving workflow.

To illustrate the closed-loop framework originating from *DigCat*, the design of a novel stable and low-cost bifunctional metal oxide (MO) electrocatalyst for water splitting was investigated.[117] From 1430 thermodynamically stable MOs identified by stability analysis in *DigCat*, $RbSbWO_6$ was chosen as a case study. $RbSbWO_6$ outperformed several widely studied, heavily engineered MOs for HER and OER under acidic media. The addition of the $RbSbWO_6$ experimental dataset to *DigCat*'s database further improves catalyst discovery iteratively, addressing the limitation of high-quality experimental datasets. This study demonstrates the closed-loop framework and the potential of digital materials platforms' AI-driven automated workflow in integrating human curiosity, theory-based knowledge and machine precision effectively to accelerate real world impact.

In a parallel development, high-entropy alloys (HEAs) have emerged as one of the most challenging material systems due to their vast compositional complexity and multi-principal element design space. The conventional trial-and-error synthesis method is inefficient for such a large design space. High-throughput and data-driven ML strategies have been systematically reviewed as key enablers for accelerated HEA

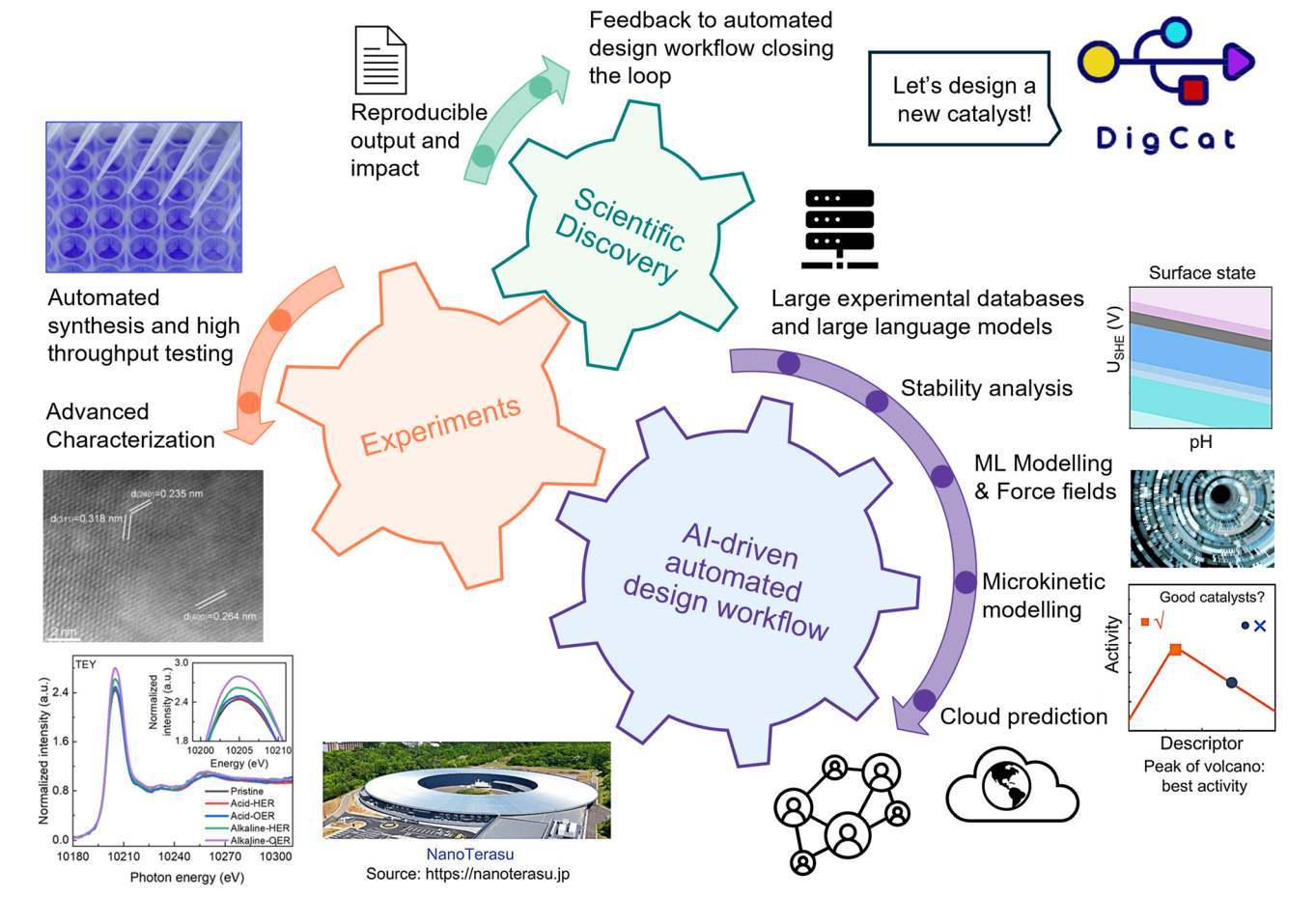

Fig. 11 Inclusion of a thorough AI-driven automated design workflow towards accelerating scientific discoveries in a closed-loop framework (reference place holder for cloud synthesis).[117] The integrated design framework connects high-throughput experiments, AI-driven automated workflows, and scientific insights to accelerate catalyst development. By coupling robotic experimentation and advanced characterization with ML-guided screening, descriptor analysis, and mechanistic understanding, the *DigMat* platform enables continuous feedback between data, models, and experiments. Adapted with permission from ref. 117 © 2024 The Authors.

discovery, covering preparation, characterization, computation, and structure–property mapping necessary for efficient exploration of HEA systems.[118] Emerging autonomous experimental platforms are advancing the integration of AI with robotics for materials synthesis. Notably, the concept of "self-driving laboratories" that integrate AI, automation, and high-throughput characterization is gaining traction in alloy development and broader materials science, showing that automated closed-loop workflows can unify predictive models with robotic experimentation to efficiently explore new compositions.[119] Even in HEA development, machine learning models are being combined with high-throughput synthesis methods to generate large libraries of alloy samples rapidly, and automated characterization data are fed back into ML models for iterative design optimization.[120] These advances highlight that autonomous and AI-guided experimental systems are increasingly important for HEA design, helping to overcome data scarcity and accelerate the translation of predictions into verified materials.

7. Perspectives and outlook

The digital materials ecosystem represents a transformative paradigm in which data, physical models, machine intelligence, AI agents, and automated experimentation operate as a unified framework for materials discovery. Although rapid progress has been made, critical challenges remain in data reliability, model interpretability, and seamless integration between digital design and physical validation. Looking ahead, establishing trusted benchmark datasets, advancing physics-informed and explainable AI models, developing reasoning-capable AI agents, and achieving standardized closed-loop automation will be essential. Future efforts should focus on the following directions.

7.1. Data reliability and benchmarking

To make digital materials truly "scientific," the underlying data must be reliable, traceable, and benchmarked against high-quality experimental standards. While massive data accumulation has been achieved, the next milestone lies in data verification—establishing universal protocols for reproducibility, metadata completeness, and inter-database consistency. High-quality benchmark datasets that connect experimental and simulated results are indispensable for model calibration, uncertainty quantification, and trustworthy AI predictions.

7.2. "Human intelligence" is important – human knowledge as the training ground for AI agents

AI agents are powerful in reasoning, yet they remain reflections of the data and scientific logic we provide. The next generation of materials AI should be trained not only on raw data but also on the thought processes of human scientists—how hypotheses are formed, how contradictory evidence is resolved, and how causality is reasoned. Encoding such cognitive patterns will enable AI agents to move beyond correlation learning toward mechanism discovery, transforming them from assistants to collaborative scientists.

7.3. Toward semantic descriptors and high-precision models

Traditional ML models rely heavily on numerical descriptors derived from structure or composition. However, the complexity of real materials often defies such simplifications. Developing semantic-based descriptors – where latent physical meaning is encoded through language and multimodal embeddings – can bridge symbolic scientific reasoning with numerical prediction. This direction promises high-precision models that understand materials not just as data points, but as semantic entities within a scientific context.

7.4. Establishing autonomous mechanistic discovery workflows

The ability of AI agents to uncover underlying mechanisms is still at an early stage. There is an urgent need for a standardized and interpretable workflow that spans data extraction, hypothesis formulation, and rigorous mechanistic validation. Embedding symbolic regression, causal inference, and uncertainty-aware exploration within agent pipelines will make mechanism discovery systematic and reproducible rather than accidental.

7.5. Standardization and alignment in closed-loop experimentation

Finally, realizing a true self-driving laboratory requires seamless communication between digital design and physical execution. This involves not only hardware–software interfacing but also semantic alignment—ensuring that the agent's "intent" matches the experimental system's "response." Defining common data schemas, input/output formats, and validation metrics will be crucial for integrating AI reasoning with high-throughput synthesis and characterization.

In summary, the future of modern digital materials will depend on a dual evolution: the scientific rigor of data and models, and the cognitive sophistication of AI agents. By merging verified data, interpretable models, human-inspired reasoning, and standardized automation, the community can move from knowledge accumulation to autonomous scientific discovery—a transition that may redefine not only materials research but the very process of scientific innovation itself.

Author contributions

Di Zhang led the manuscript preparation, including overall writing, structuring the framework, compiling and integrating contributions from all co-authors, and organizing the figures. Xue Jia wrote the machine learning model section. Yuhang Wang wrote the database-related section. Heng Liu wrote the catalysis theoretical modeling section. Qian Wang wrote the battery materials section. Seong-Hoon Jang wrote the hydrogen storage materials section. Daksh Shah wrote the closed-loop experimental process/workflow section. Songbo Ye prepared the visualizations/figures for the digital materials ecosystem. Hung Ba Tran wrote the theoretical modeling section for hydrogen storage. Hao Li proposed the concept of digital materials

ecosystem and the overall idea of the work, and supervised and coordinated the project. All authors discussed the results and contributed to revising the manuscript.

Conflicts of interest

There are no conflicts to declare.

Data availability

No primary research results, software or code have been included and no new data were generated or analysed as part of this perspective.

Acknowledgements

We acknowledge support from JSPS KAKENHI (No. JP25K01737, JP25K17991, and JP24K23068).

References

1 L. Himanen, A. Geurts, A. S. Foster and P. Rinke, Data-Driven Materials Science: Status, Challenges, and Perspectives, *Adv. Sci.*, 2019, **6**(21), 1900808.
2 F. Zivic, A. K. Malisic, N. Grujovic, B. Stojanovic and M. Ivanovic, Materials Informatics: A Review of Ai and Machine Learning Tools, Platforms, Data Repositories, and Applications to Architectured Porous Materials, *Mater. Today Commun.*, 2025, **48**, 113525.
3 M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C.-T. Dinh, P. De Luna, Z. Yu, A. S. Rasouli, P. Brodersen, *et al.*, Accelerated Discovery of Co2 Electrocatalysts Using Active Machine Learning, *Nature*, 2020, **581**(7807), 178–183.
4 J. A. Esterhuizen, B. R. Goldsmith and S. Linic, Interpretable Machine Learning for Knowledge Generation in Heterogeneous Catalysis, *Nat. Catal.*, 2022, **5**(3), 175–184.
5 N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, *et al.*, An Autonomous Laboratory for the Accelerated Synthesis of Novel Materials, *Nature*, 2023, **624**(7990), 86–91.
6 Z. Di, J. Xue, L. Heng, W. Yuhang, Y. Songbo, J. Qiuling, W. Yuan, G. Zhongyuan, Z. Linda, W. Li, *et al.*, Cloud Synthesis: A Global Close-Loop Feedback Powered by Autonomous Ai-Driven Catalyst Design Agent, *AI Agent*, 2025, **1**(1), 2.
7 E. I. Marchenko, S. A. Fateev, A. A. Petrov, V. V. Korolev, A. Mitrofanov, A. V. Petrov, E. A. Goodilin and A. B. Tarasov, Database of Two-Dimensional Hybrid Perovskite Materials: Open-Access Collection of Crystal Structures, Band Gaps, and Atomic Partial Charges Predicted by Machine Learning, *Chem. Mater.*, 2020, **32**(17), 7383–7388.
8 L. Sarkisov, R. Bueno-Perez, M. Sutharson and D. Fairen-Jimenez, Materials Informatics with Poreblazer V4.0 and the Csd Mof Database, *Chem. Mater.*, 2020, **32**(23), 9849–9867.
9 M. de Jong, W. Chen, H. Geerlings, M. Asta and K. A. Persson, A Database to Enable Discovery and Design of Piezoelectric Materials, *Sci. Data*, 2015, **2**, 150053.
10 F. Ricci, W. Chen, U. Aydemir, G. J. Snyder, G. M. Rignanese, A. Jain and G. Hautier, An *Ab Initio* Electronic Transport Database for Inorganic Materials, *Sci. Data*, 2017, **4**, 170085.
11 F. A. Rasmussen and K. S. Thygesen, Computational 2d Materials Database: Electronic Structure of Transition-Metal Dichalcogenides and Oxides, *J. Phys. Chem. C*, 2015, **119**(23), 13169–13183.
12 L. Ward, S. Babinec, E. J. Dufek, D. A. Howey, V. Viswanathan, M. Aykol, D. A. C. Beck, B. Blaiszik, B.-R. Chen, G. Crabtree, *et al.*, Principles of the Battery Data Genome, *Joule*, 2022, **6**(10), 2253–2271.
13 A. Belsky, M. Hellenbrandt, V. L. Karen and P. Luksch, New Developments in the Inorganic Crystal Structure Database (Icsd): Accessibility in Support of Materials Research and Design, *Acta Crystallogr., Sect. B*, 2002, **58**(1), 364–369.
14 A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. White, K. Munch, W. Tumas and C. Phillips, An Open Experimental Database for Exploring Inorganic Materials, *Sci. Data*, 2018, **5**, 180053.
15 X. Qu, A. Jain, N. N. Rajput, L. Cheng, Y. Zhang, S. P. Ong, M. Brafman, E. Maginn, L. A. Curtiss and K. A. Persson, The Electrolyte Genome Project: A Big Data Approach in Battery Materials Discovery, *Comput. Mater. Sci.*, 2015, **103**, 56–67.
16 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, *et al.*, Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation, *APL Mater.*, 2013, (1), 011002.
17 S. S. Borysov, R. M. Geilhufe and A. V. Balatsky, Organic Materials Database: An Open-Access Online Database for Data Mining, *PLoS One*, 2017, **12**(2), e0171501.
18 S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, *et al.*, The Computational 2d Materials Database: High-Throughput Modeling and Discovery of Atomically Thin Crystals, *2D Materials*, 2018, **5**(4), 042002.
19 M. N. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, F. Bertoldo, T. Deilmann, N. R. Knøsgaard, M. Kruse, A. H. Larsen, S. Manti, *et al.*, Recent Progress of the Computational 2d Materials Database (C2db), *2D Materials*, 2021, **8**(4), 044002.
20 S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl and C. Wolverton, The Open Quantum Materials Database (Oqmd): Assessing the Accuracy of Dft Formation Energies, *npj Comput. Mater.*, 2015, **1**, 15010.
21 J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (Oqmd), *JOM*, 2013, **65**(11), 1501–1509.
22 J. Hu, S. Stefanov, Y. Song, S. S. Omee, S.-Y. Louis, E. M. D. Siriwardane, Y. Zhao and L. Wei, Materialsatlas.Org: A Materials Informatics Web App

Platform for Materials Discovery and Survey of State-of-the-Art, *npj Comput. Mater.*, 2022, **8**, 65.

23 K. T. Winther, M. J. Hoffmann, J. R. Boes, O. Mamun, M. Bajdich and T. Bligaard, Catalysis-Hub.Org, an Open Electronic Structure Database for Surface Reactions, *Sci. Data*, 2019, **6**(1), 75.

24 J. Zhou, L. Shen, M. D. Costa, K. A. Persson, S. P. Ong, P. Huck, Y. Lu, X. Ma, Y. Chen, H. Tang, *et al.*, 2dmatpedia, an Open Computational Database of Two-Dimensional Materials from Top-down and Bottom-up Approaches, *Sci. Data*, 2019, **6**(1), 86.

25 S. Huang and J. M. Cole, A Database of Battery Materials Auto-Generated Using Chemdataextractor, *Sci. Data*, 2020, **7**(1), 260.

26 L. Sbailò, Á. Fekete, L. M. Ghiringhelli and M. Scheffler, The Nomad Artificial-Intelligence Toolkit: Turning Materials-Science Data into Knowledge and Understanding, *npj Comput. Mater.*, 2022, **8**(1), 250.

27 S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, *et al.*, Aflow: An Automatic Framework for High-Throughput Materials Discovery, *Comput. Mater. Sci.*, 2012, **58**, 218–226.

28 M. Huang, R. Shi, H. Liu, W. Ding, J. Fan, B. Zhou, B. Da, Z. Gao, H. Li and W. Yang, Computational Single-Atom Catalyst Database Empowers the Machine Learning Assisted Design of High-Performance Catalysts, *J. Phys. Chem. C*, 2025, **129**(10), 5043–5053.

29 D. Zhang and H. Li, *Digital Catalysis Platform (Digcat): A Gateway to Big Data and Ai-Powered Innovations in Catalysis*, 2024.

30 F. Yang, E. Campos dos Santos, X. Jia, R. Sato, K. Kisu, Y. Hashimoto, S.-i. Orimo and H. Li, A Dynamic Database of Solid-State Electrolyte (Ddse) Picturing All-Solid-State Batteries, *Nano Mater. Sci.*, 2024, **6**(2), 256–262.

31 D. Zhang, X. Jia, H. B. Tran, S. H. Jang, L. Zhang, R. Sato, Y. Hashimoto, T. Sato, K. Konno, S.-i. Orimo, *et al.*, "DIVE" into Hydrogen Storage Materials Discovery with AI Agents, *Chem. Sci.*, 2026, **17**, 3031–3042.

32 J. Abed, J. Kim, M. Shuaibi, B. Wander, B. Duijf, S. Mahesh, H. Lee, V. Gharakhanyan, S. Hoogland and E. Irtem, Open Catalyst Experiments 2024 (Ocx24): Bridging Experiments and Computational Models, *arXiv*, 2024, preprint, arXiv:2411.11783, DOI: 10.48550/arXiv.2411.11783.

33 D. Sivan, K. Satheesh Kumar, A. Abdullah, V. Raj, I. I. Misnon, S. Ramakrishna and R. Jose, Advances in Materials Informatics: A Review, *J. Mater. Sci.*, 2024, **59**(7), 2602–2643.

34 R. Mahajan, A. M. Aleman, C. F. Crago, S. Bhasker-Ranganath, M. E. Kreider, J. A. Zamora Zeledon, J. Schröder, G. A. Kamat, M. A. Hubert, A. C. Nielander, *et al.*, A Research Database for Experimental Electrocatalysis: Advancing Data Sharing and Reusability, *J. Chem. Phys.*, 2025, **163**, 124704.

35 H. A. Hansen, J. Rossmeisl and J. K. Nørskov, Surface Pourbaix Diagrams and Oxygen Reduction Activity of Pt, Ag and Ni (111) Surfaces Studied by Dft, *Phys. Chem. Chem. Phys.*, 2008, **10**(25), 3722–3730.

36 H. Liu, X. Jia, A. Cao, L. Wei, C. D'agostino and H. Li, The Surface States of Transition Metal X-Ides under Electrocatalytic Conditions, *J. Chem. Phys.*, 2023, **158**, 124705.

37 H. Liu, D. Zhang, Y. Wang and H. Li, Reversible Hydrogen Electrode (Rhe) Scale Dependent Surface Pourbaix Diagram at Different Ph, *Langmuir*, 2024, **40**(14), 7632–7638.

38 H. Fei, J. Dong, Y. Feng, C. S. Allen, C. Wan, B. Volosskiy, M. Li, Z. Zhao, Y. Wang and H. Sun, General Synthesis and Definitive Structural Identification of Mn4c4 Single-Atom Catalysts with Tunable Electrocatalytic Activities, *Nat. Catal.*, 2018, **1**(1), 63–72.

39 J. K. Nørskov, J. Rossmeisl, A. Logadottir, L. Lindqvist, J. R. Kitchin, T. Bligaard and H. Jonsson, Origin of the Overpotential for Oxygen Reduction at a Fuel-Cell Cathode, *J. Phys. Chem. B*, 2004, **108**(46), 17886–17892.

40 G. Henkelman, B. P. Uberuaga and H. Jónsson, A Climbing Image Nudged Elastic Band Method for Finding Saddle Points and Minimum Energy Paths, *J. Chem. Phys.*, 2000, **113**(22), 9901–9904.

41 G. Henkelman and H. Jónsson, Improved Tangent Estimate in the Nudged Elastic Band Method for Finding Minimum Energy Paths and Saddle Points, *J. Chem. Phys.*, 2000, **113**(22), 9978–9985.

42 S. R. Kelly, C. Kirk, K. Chan and J. K. Nørskov, Electric Field Effects in Oxygen Reduction Kinetics: Rationalizing Ph Dependence at the Pt (111), Au (111), and Au (100) Electrodes, *J. Phys. Chem. C*, 2020, **124**(27), 14581–14591.

43 H. Li, S. Kelly, D. Guevarra, Z. Wang, Y. Wang, J. A. Haber, M. Anand, G. K. K. Gunasooriya, C. S. Abraham and S. Vijay, Analysis of the Limitations in the Oxygen Reduction Activity of Transition Metal Oxide Surfaces, *Nat. Catal.*, 2021, **4**(6), 463–468.

44 D. Zhang, Y. Hirai, K. Nakamura, K. Ito, Y. Matsuo, K. Ishibashi, Y. Hashimoto, H. Yabu and H. Li, Benchmarking Ph-Field Coupled Microkinetic Modeling against Oxygen Reduction in Large-Scale Fe–Azaphthalocyanine Catalysts, *Chem. Sci.*, 2024, **15**(14), 5123–5132.

45 D. Zhang, Z. Wang, F. Liu, P. Yi, L. Peng, Y. Chen, L. Wei and H. Li, Unraveling the Ph-Dependent Oxygen Reduction Performance on Single-Atom Catalysts: From Single- to Dual-Sabatier Optima, *J. Am. Chem. Soc.*, 2024, **146**(5), 3210–3219.

46 D. Zhang, F. She, J. Chen, L. Wei and H. Li, Why Do Weak-Binding M–N–C Single-Atom Catalysts Possess Anomalously High Oxygen Reduction Activity?, *J. Am. Chem. Soc.*, 2025, **147**(7), 6076–6086.

47 Q. Jiang, M. Gu, S. Pei, T. Wang, F. Liu, X. Yang, D. Zhang, Z. Wu, Y. Wang and L. Wei, The Key Steps and Distinct Performance Trends of Pyrrolic *vs.* Pyridinic M–N–C Catalysts in Electrocatalytic Nitrate Reduction, *J. Am. Chem. Soc.*, 2025, **147**(29), 26029–26039.

48 Y. Wang, D. Zhang, B. Sun, X. Jia, L. Zhang, H. Cheng, J. Fan and H. Li, Divergent Activity Shifts of Tin-Based Catalysts for Electrochemical Co2 Reduction: Ph-Dependent Behavior of Single-Atom *Versus* Polyatomic Structures, *Angew. Chem., Int. Ed.*, 2025, **64**(8), e202418228.
49 Y.-C. Gao, X. Chen, Y.-H. Yuan, Y.-P. Chen, Y.-L. Niu, N. Yao, Y.-B. Gao, W.-L. Li and Q. Zhang, Accelerating Battery Innovation: Ai-Powered Molecular Discovery, *Chem. Soc. Rev.*, 2025, **54**, 9630–9684.
50 E. Campos dos Santos, R. Sato, K. Kisu, K. Sau, X. Jia, F. Yang, S.-i. Orimo and H. Li, Explore the Ionic Conductivity Trends on B12h12 Divalent Closo-Type Complex Hydride Electrolytes, *Chem. Mater.*, 2023, **35**(15), 5996–6004.
51 Q. Wang, F. Yang, Y. Wang, D. Zhang, R. Sato, L. Zhang, E. J. Cheng, Y. Yan, Y. Chen, K. Kisu, *et al.*, Unraveling the Complexity of Divalent Hydride Electrolytes in Solid-State Batteries *via* a Data-Driven Framework with Large Language Model, *Angew. Chem., Int. Ed.*, 2025, **64**(25), e202506573.
52 F. J. DiSalvo, Thermoelectric Cooling and Power Generation, *Science*, 1999, **285**(5428), 703–706.
53 X. Jia, S. Li, Z. Zhang, Y. Deng, X. Li, Y. Cao, Y. Yan, J. Mao, J. Yang, Q. Zhang, *et al.*, Using Materials Quality Factor Bδε* for Design of Thermoelectric Materials with Multiple Bands, *Mater. Today Phys.*, 2021, **18**, 100371.
54 Y. Iwasaki, I. Takeuchi, V. Stanev, A. G. Kusne, M. Ishida, A. Kirihara, K. Ihara, R. Sawada, K. Terashima, H. Someya, *et al.*, Machine-Learning Guided Discovery of a New Thermoelectric Material, *Sci. Rep.*, 2019, **9**(1), 2751.
55 A. Furmanchuk, J. E. Saal, J. W. Doak, G. B. Olson, A. Choudhary and A. Agrawal, Prediction of Seebeck Coefficient for Compounds without Restriction to Fixed Stoichiometry: A Machine Learning Approach, *J. Comput. Chem.*, 2018, **39**(4), 191–201.
56 H. Yuan, S. Han, R. Hu, W. Y. Jiao, M. Li, H. Liu and Y. Fang, Machine Learning for Accelerated Prediction of the Seebeck Coefficient at Arbitrary Carrier Concentration, *Mater. Today Phys.*, 2022, **25**, 100706.
57 I. Ullah, K. Ullah, L. F. Zhao and Z. F. Zhou, Machine Learning-Driven Multiobjective Optimization of a Mems in-Situ Thermoelectric Seebeck Coefficient Measurement Structure, *Measurement*, 2025, **256**, 118500.
58 A. L. Ben Kamri, M. A. Fadla, I. k. Lefkaier, C. l. Ben Messaoud, M. B. Kanoun and S. Goumri-Said, Ai-Driven Ensemble Learning for Accurate Seebeck Coefficient Prediction in Half-Heusler Compounds Based on Chemical Formulas, *Comput. Condens. Matter*, 2024, **40**, e00923.
59 T. A. Alrebdi, Y. S. Wudil, U. F. Ahmad, F. A. Yakasai, J. Mohammed and F. H. Alkallas, Predicting the Thermal Conductivity of BiTe-Based Thermoelectric Energy Materials: A Machine Learning Approach, *Int. J. Therm. Sci.*, 2022, **181**, 107784.
60 Q. Ren, D. Chen, L. Rao, Y. Lun, G. Tang and J. Hong, Machine-Learning-Assisted Discovery of 212-Zintl-Phase Compounds with Ultra-Low Lattice Thermal Conductivity, *J. Mater. Chem. A*, 2024, **12**(2), 1157–1165.
61 S. Zeng, L. Fang, Z. Gu, X. Wang, Y. Zhao, G. Li, Y. Tu and J. Ni, Ultralow Lattice Thermal Conductivities and Excellent Thermoelectric Properties of Hypervalent Triiodides Xi3 (X = Rb, Cs) Discovered by Machine Learning Method, *J. Chem. Phys.*, 2023, **159**, 014703.
62 T. S. Zhu, R. He, S. Gong, T. Xie, P. Gorai, K. Nielsch and J. C. Grossman, Charting Lattice Thermal Conductivity for Inorganic Crystals and Discovering Rare Earth Chalcogenides for Thermoelectrics, *Energ Environ. Sci.*, 2021, **14**(6), 3559–3566.
63 Y. X. Zeng, W. Cao, T. Peng, Y. Hou, L. Miao, Z. Y. Wang and J. Shi, A Machine Learning-Based Framework for Predicting the Power Factor of Thermoelectric Materials, *Appl. Mater. Today*, 2025, **43**, 102627.
64 Y. Sheng, Y. S. Wu, J. Yang, W. C. Lu, P. Villars and W. Q. Zhang, Active Learning for the Power Factor Prediction in Diamond-Like Thermoelectric Materials, *npj Comput. Mater.*, 2020, **6**(1), 171.
65 Z. Yang, Y. Sheng, C. Zhu, J. Y. Ni, Z. Y. Zhu, J. Y. Xi, W. Zhang and J. Yang, Accurate and Explainable Machine Learning for the Power Factors of Diamond-Like Thermoelectric Materials, *J Materiomics*, 2022, **8**(3), 633–639.
66 Y. Wang, C. Zhong, J. Zhang, H. Yao, J. Chen and X. Lin, High-Performance Stacking Ensemble Learning for Thermoelectric Figure-of-Merit Prediction, *Mater. Des.*, 2025, **249**, 113552.
67 X. Jia, A. Aziz, Y. Hashimoto and H. Li, Dealing with the Big Data Challenges in Ai for Thermoelectric Materials, *Sci. China Mater.*, 2024, **67**(4), 1173–1182.
68 Y. Gan, G. J. Wang, J. Zhou and Z. M. Sun, Prediction of Thermoelectric Performance for Layered Iv-V-Vi Semiconductors by High-Throughput *Ab Initio* Calculations and Machine Learning, *npj Comput. Mater.*, 2021, **7**(1), 176.
69 C. T. Ma and S. J. Poon, Reexamining Machine Learning Models on Predicting Thermoelectric Properties, *arXiv*, 2025, preprint, arXiv:2509.00299, DOI: 10.48550/arXiv.2509.00299.
70 D. Chernyavsky, J. Van den Brink, G. H. Park, K. Nielsch and A. Thomas, Sustainable Thermoelectric Materials Predicted by Machine Learning, *Adv. Theory Simul.*, 2022, **5**(11), 2200351.
71 X. Jia, Y. S. Deng, X. Bao, H. H. Yao, S. Li, Z. Li, C. Chen, X. Y. Wang, J. Mao, F. Cao, *et al.*, Unsupervised Machine Learning for Discovery of Promising Half-Heusler Thermoelectric Materials, *npj Comput. Mater.*, 2022, **8**(1), 34.
72 V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder and A. J. N. Jain, Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature, *Nature*, 2019, **571**(7763), 95–98.
73 X. Jia, H. H. Yao, Z. J. Yang, J. Y. Shi, J. X. Yu, R. P. Shi, H. J. Zhang, F. Cao, X. Lin, J. Mao, *et al.*, Advancing

Thermoelectric Materials Discovery through Semi-Supervised Learning and High-Throughput Calculations, *Appl. Phys. Lett.*, 2023, **123**, 203902.

74 Y. Long, C. Zhong, X. Ma, J. Zhang, H. Yao, J. Liu, K. Hu, Q. Zhang and X. Lin, Inverse Design of High-Performance Thermoelectric Materials *via* a Generative Model Combined with Experimental Verification, *ACS Appl. Mater. Interfaces*, 2025, **17**(13), 19856–19867.

75 L. M. Antunes, Vikram, J. J. Plata, A. V. Powell, K. T. Butler and R. Grau-Crespo, Machine Learning Approaches for Accelerating the Discovery of Thermoelectric Materials, in *Machine Learning in Materials Informatics: Methods and Applications, Acs Symposium Series*, vol. 1416, American Chemical Society, 2022, pp. 1–32.

76 W. Yuelin, Z. Chengquan, Z. Jingzi, L. Jiakai, H. Kailong, C. Junjie and L. Xi, Machine Learning for Predictive Design and Optimization of High-Performance Thermoelectric Materials: A Review, *J. Mater. Inform.*, 2025, **5**(3), 41.

77 Y. Wu, D. Song, M. An, C. Chi, C. Zhao, B. Yao, W. Ma and X. Zhang, Unlocking New Possibilities in Ionic Thermoelectric Materials: A Machine Learning Perspective, *Natl. Sci. Rev.*, 2025, **12**(1), nwae411.

78 S. N. Steinmann, Q. Wang and Z. W. Seh, How Machine Learning Can Accelerate Electrocatalysis Discovery and Optimization, *Mater. Horiz.*, 2023, **10**(2), 393–406.

79 C. Pitchai, T. Y. Lo, H. C. Chang, H. C. Li, M. D. Yang and C. M. Chen, Machine Learning-Assisted Optimization Design for Enhanced Oxygen Evolution Reaction Based on Vanadium-Doped Nickel-Cobalt Layered Double Hydroxides, *J. Mater. Chem. A*, 2025, **13**, 28907–28919.

80 X. Jiang, Y. Wang, B. Jia, X. Qu and M. Qin, Prediction of Oxygen Evolution Activity for Nicofe Oxide Catalysts *Via* Machine Learning, *ACS Omega*, 2022, **7**(16), 14160–14164.

81 X. Jia and H. Li, Machine Learning Enabled Exploration of Multicomponent Metal Oxides for Catalyzing Oxygen Reduction in Alkaline Media, *J. Mater. Chem. A*, 2024, **12**(21), 12487–12500.

82 X. Wan, Z. Zhang, W. Yu, H. Niu, X. Wang and Y. Guo, Machine-Learning-Assisted Discovery of Highly Efficient High-Entropy Alloy Catalysts for the Oxygen Reduction Reaction, *Patterns*, 2022, **3**(9), 100553.

83 T. T. Shi, G. Y. Liu and Z. X. Chen, Machine Learning Prediction of Co Adsorption Energies and Properties of Layered Alloys Using an Improved Feature Selection Algorithm, *J. Phys. Chem. C*, 2023, **127**(20), 9573–9583.

84 H. Li, X. Li, P. Wang, Z. Zhang, K. Davey, J. Q. Shi and S. Z. Qiao, Machine Learning Big Data Set Analysis Reveals C-C Electro-Coupling Mechanism, *J. Am. Chem. Soc.*, 2024, **146**(32), 22850–22858.

85 J. Li, N. Wu, J. Zhang, H.-H. Wu, K. Pan, Y. Wang, G. Liu, X. Liu, Z. Yao and Q. Zhang, Machine Learning-Assisted Low-Dimensional Electrocatalysts Design for Hydrogen Evolution Reaction, *Nanomicro Lett.*, 2023, **15**(1), 227.

86 A. Kolluru, M. Shuaibi, A. Palizhati, N. Shoghi, A. Das, B. Wood, C. L. Zitnick, J. R. Kitchin and Z. W. Ulissi, Open Challenges in Developing Generalizable Large-Scale Machine-Learning Models for Catalyst Discovery, *ACS Catal.*, 2022, **12**(14), 8572–8581.

87 K. Abdelmaqsoud, M. Shuaibi, A. Kolluru, R. Cheula and J. R. Kitchin, Investigating the Error Imbalance of Large-Scale Machine Learning Potentials in Catalysis, *Catal. Sci. Technol.*, 2024, **14**(20), 5899–5908.

88 J. Lan, A. Palizhati, M. Shuaibi, B. M. Wood, B. Wander, A. Das, M. Uyttendaele, C. L. Zitnick and Z. W. Ulissi, Adsorbml: A Leap in Efficiency for Adsorption Energy Calculations Using Generalizable Machine Learning Potentials, *npj Comput. Mater.*, 2023, **9**(1), 172.

89 D. Zhang, P. Yi, X. Lai, L. Peng and H. Li, Active Machine Learning Model for the Dynamic Simulation and Growth Mechanisms of Carbon on Metal Surface, *Nat. Commun.*, 2024, **15**(1), 344.

90 A. L. Maulana, S. Han, Y. Shan, P. C. Chen, C. Lizandara-Pueyo, S. De, K. Schierle-Arndt and P. Yang, Stabilizing Ru in Multicomponent Alloy as Acidic Oxygen Evolution Catalysts with Machine Learning-Enabled Structural Insights and Screening, *J. Am. Chem. Soc.*, 2025, **147**(12), 10268–10278.

91 Y. Zeng, J. Wang, F. Li, T. Liu and A. Xu, Ai-Accelerated Discovery of Electrocatalyst Materials, *ACS Mater. Au*, 2025, **6**(1), 72–89.

92 A. I. Osman, M. Nasr, A. S. Eltaweil, M. Hosny, M. Farghali, A. S. Al-Fatesh, D. W. Rooney and E. M. Abd El-Monaem, Advances in Hydrogen Storage Materials: Harnessing Innovative Technology, from Machine Learning to Computational Chemistry, for Energy Storage Solutions, *Int. J. Hydrogen Energy*, 2024, **67**, 1270–1294.

93 T. Liu, L. Xue, B. Cheng, Y. Zhao, B. Dou, M. Paredes and L. Song, Machine Learning-Driven Alloy Digital Design for Hydrogen Storage: A Review, *Digital Twin*, 2025, **2**(3), 2511889.

94 S.-H. Jang, D. Zhang, H. B. Tran, X. Jia, K. Konno, R. Sato, S.-i. Orimo and H. Li, Physically Interpretable Descriptors Drive the Materials Design of Metal Hydrides for Hydrogen Storage, *Chem. Sci.*, 2025, **16**(48), 23111–23120.

95 U. D. O. Energy, *Targets for Onboard Hydrogen Storage Systems for Light-Duty Vehicles*, 2009.

96 M. C. Ramos, C. J. Collison and A. D. White, A Review of Large Language Models and Autonomous Agents in Chemistry, *Chem. Sci.*, 2025, **16**(6), 2514–2572.

97 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**(1), 79.

98 M. Ansari and S. M. Moosavi, Agent-Based Learning of Materials Datasets from the Scientific Literature, *Digital Discov.*, 2024, **3**(12), 2607–2617.

99 R. Odobesku, K. Romanova, S. Mirzaeva, O. Zagorulko, R. Sim, R. Khakimullin, J. Razlivina, A. Dmitrenko and V. Vinogradov, Agent-Based Multimodal Information Extraction for Nanomaterials, *npj Comput. Mater.*, 2025, **11**(1), 194.

100 F. Yang, E. Campos dos Santos, X. Jia, R. Sato, K. Kisu, Y. Hashimoto, S.-i. Orimo and H. Li, A Dynamic Database

of Solid-State Electrolyte (Ddse) Picturing All-Solid-State Batteries, *Nano Mater. Sci.*, 2024, **6**(2), 256–262.

101 F.-L. Yang, R. Sato, E. J.-F. Cheng, K. Kisu, Q. Wang, X. Jia, S.-i. Orimo and H. Li, Data-Driven Viewpoint for Developing Next-Generation Mg-Ion Solid-State Electrolytes, *J. Electrochem.*, 2024, **30**(7), 3.

102 Z. Zhang, J. Chu, H. Zhang, X. Liu and M. He, Mining Ionic Conductivity Descriptors of Antiperovskite Electrolytes for All-Solid-State Batteries *Via* Machine Learning, *J. Energy Storage*, 2024, **75**, 109714.

103 H. A. Cortés, M. R. Bonilla, H. Früchtl, T. van Mourik, J. Carrasco and E. A. Akhmatskaya, Data-Mining Approach to Understanding the Impact of Multi-Doping on the Ionic Transport Mechanism of Solid Electrolytes Materials: The Case of Dual-Doped $Ga_{0.15}/Sc_yli_7la_3zr_2o_{12}$, *J. Mater. Chem. A*, 2024, **12**(9), 5181–5193.

104 H. J. Lee, H. Kim, S. Ji, K. Choi, H. Choi, W. Lim and B. Lee, Lithium Localization by Anions in Argyrodite Solid Electrolytes from Machine-Learning-Based Simulations, *Adv. Energy Mater.*, 2024, **14**(48), 2402396.

105 G. Bradford, J. Lopez, J. Ruza, M. A. Stolberg, R. Osterude, J. A. Johnson, R. Gomez-Bombarelli and Y. Shao-Horn, Chemistry-Informed Machine Learning for Polymer Electrolyte Discovery, *ACS Cent. Sci.*, 2023, **9**(2), 206–216.

106 B. Liu, J. Yang, H. Yang, C. Ye, Y. Mao, J. Wang, S. Shi, J. Yang and W. Zhang, Rationalizing the Interphase Stability of Li|Doped-$Li_7la_3zr_2o_{12}$ *Via* Automated Reaction Screening and Machine Learning, *J. Mater. Chem. A*, 2019, **7**(34), 19961–19969.

107 H. Hwang, H. Jeong, J.-W. Cho, Y. Oh, D. Kim, D. Shin, J.-H. Lee, H. Kim and J.-H. Hwang, Machine Learning-Assisted Microstructural Quantification of Multiphase Cathode Composites in All-Solid-State Batteries: Correlation with Battery Performance, *Small*, 2025, **21**(10), 2410016.

108 J. Bai, S. Mosbach, C. J. Taylor, D. Karan, K. F. Lee, S. D. Rihm, J. Akroyd, A. A. Lapkin and M. Kraft, A Dynamic Knowledge Graph Approach to Distributed Self-Driving Laboratories, *Nat. Commun.*, 2024, **15**(1), 462.

109 D. A. Boiko, R. MacKnight, B. Kline and G. Gomes, Autonomous Chemical Research with Large Language Models, *Nature*, 2023, **624**(7992), 570–578.

110 N. Yoshikawa, M. Skreta, K. Darvish, S. Arellano-Rubach, Z. Ji, L. Bjørn Kristensen, A. Z. Li, Y. Zhao, H. Xu, A. Kuramshin, *et al.*, Large Language Models for Chemistry Robotics, *Autonomous Robots*, 2023, **47**(8), 1057–1086.

111 D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, *et al.*, Accelerating the Discovery of Materials for Clean Energy in the Era of Smart Automation, *Nat. Rev. Mater.*, 2018, **3**(5), 5–20.

112 G. Tom, S. P. Schmid, S. G. Baird, Y. Cao, K. Darvish, H. Hao, S. Lo, S. Pablo-García, E. M. Rajaonson, M. Skreta, *et al.*, Self-Driving Laboratories for Chemistry and Materials Science, *Chem. Rev.*, 2024, **124**(16), 9633–9732.

113 E. Stach, B. DeCost, A. G. Kusne, J. Hattrick-Simpers, K. A. Brown, K. G. Reyes, J. Schrier, S. Billinge, T. Buonassisi, I. Foster, *et al.*, Autonomous Experimentation Systems for Materials Development: A Community Perspective, *Matter*, 2021, **4**(9), 2702–2726.

114 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 Discov.*, 2024, **3**(3), 491–501.

115 Y. Ruan, C. Lu, N. Xu, Y. He, Y. Chen, J. Zhang, J. Xuan, J. Pan, Q. Fang, H. Gao, *et al.*, An Automatic End-to-End Chemical Synthesis Development Platform Powered by Large Language Models, *Nat. Commun.*, 2024, **15**(1), 10160.

116 A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon and E. D. Cubuk, Scaling Deep Learning for Materials Discovery, *Nature*, 2023, **624**(7990), 80–85.

117 X. Jia, Z. Zhou, F. Liu, T. Wang, Y. Wang, D. Zhang, H. Liu, Y. Wang, S. Ye, K. Amezawa, *et al.*, Closed-Loop Framework for Discovering Stable and Low-Cost Bifunctional Metal Oxide Catalysts for Efficient Electrocatalytic Water Splitting in Acid, *J. Am. Chem. Soc.*, 2025, **147**(26), 22642–22654.

118 L. Zhichao, M. Dong, L. Xiongjun and Z. Lu, High-Throughput and Data-Driven Machine Learning Techniques for Discovering High-Entropy Alloys, *Commun. Mater.*, 2024, **5**(1), 76.

119 B. L. Boyce and M. D. Uchic, Progress toward Autonomous Experimental Systems for Alloy Development, *MRS Bull.*, 2019, **44**(4), 273–280.

120 P. Nelaturu, J. R. Hattrick-Simpers, M. Moorehead, V. Jambur, I. Szlufarska, A. Couet and D. J. Thoma, Multi-Principal Element Alloy Discovery Using Directed Energy Deposition and Machine Learning, *Mater. Sci. Eng., A*, 2024, **891**, 145945.