



COMMENTARY

From Data to Discovery: How AI-Driven Materials Databases Are Reshaping Research

Yaping Qi^{1,*} and Weijie Yang^{2,*}

¹Advanced Institute for Materials Research (WPI-AIMR), Tohoku University, Sendai, 980-8577, Japan

²Department of Power Engineering, North China Electric Power University, Baoding, 071003, China

*Corresponding Authors: Yaping Qi. Email: qi.yaping.a2@tohoku.ac.jp; Weijie Yang. Email: yangwj@ncepu.edu.cn

Received: 03 February 2025; Accepted: 18 March 2025; Published: 16 April 2025

ABSTRACT: AI-driven materials databases are transforming research by integrating experimental and computational data to enhance discovery and optimization. Platforms such as Digital Catalysis Platform (DigCat) and Dynamic Database of Solid-State Electrolyte (DDSE) demonstrate how machine learning and predictive modeling can improve catalyst and solid-state electrolyte development. These databases facilitate data standardization, high-throughput screening, and cross-disciplinary collaboration, addressing key challenges in materials informatics. As AI techniques advance, materials databases are expected to play an increasingly vital role in accelerating research and innovation.

KEYWORDS: Data-driven; materials database; AI assistant; materials design

1 Introduction

The rapid evolution of materials science has been significantly influenced by the integration of data-driven methodologies. Traditional approaches, relying on trial-and-error experimentation and computational modeling, have faced challenges such as data fragmentation, high costs, and time-consuming validation processes. To overcome these limitations, AI-driven materials databases are emerging as transformative tools, revolutionizing how materials are discovered, analyzed, and optimized.

Recent advancements in materials informatics have demonstrated the potential of large-scale data repositories to accelerate materials discovery. A variety of materials databases, ranging from computational repositories such as Materials Project [1], ICSD (Inorganic Crystal Structure Database) [2], and Aflowlib [3], to emerging AI-enhanced platforms like DigCat [4] and DDSE [5,6], are providing valuable resources to the scientific community. While each of these databases has unique strengths, the integration of experimental and computational data, coupled with AI-driven analytics, represents a major step forward in addressing challenges such as data standardization, predictive modeling, and real-world validation.

2 The Role of Materials Databases in Scientific Research

Materials databases have evolved to serve a broad range of applications in computational and experimental materials science. These repositories enable researchers to:

- **Access structured materials data:** Consolidating experimental measurements and computational predictions in a centralized format.



- **Data mining to search materials with target function:** Based on the database, users can find materials with target functions for specific applications. Notable examples include the search of new stable metal oxide materials for electrocatalysis [7–9].
- **Facilitate high-throughput screening:** Leveraging computational modeling to predict new materials with desirable properties.
- **Support AI-driven insights:** Utilizing machine learning (ML) and large language models (LLMs) to extract patterns from vast datasets.
- **Enhance cross-disciplinary research:** Bridging experimental and theoretical studies to accelerate material discovery cycles.

However, key challenges remain in optimizing database interoperability, integrating diverse data sources, and ensuring consistency in reported experimental and computational results. Addressing these challenges requires continued innovation in AI-driven platforms that merge predictive modeling with experimental validation.

3 The Innovations of DigCat and DDSE

3.1 DigCat: The First AI-Driven Experimental Catalysis Database

The Digital Catalysis Platform (DigCat: <https://www.digcat.org/>) is a pioneering AI-powered database integrating over 800,000 experimental and computational catalyst data points [7]. Unlike traditional repositories that primarily house DFT-calculated properties, DigCat combines real experimental results with AI-driven predictive models, providing a closed-loop feedback mechanism to accelerate catalyst discovery.

Key features of DigCat:

- **Dynamic visualization tools:** Allows researchers to interact with high-dimensional data for better trend analysis.
- **LLM-based literature mining:** Extracts key insights from scientific publications to expand the database dynamically.
- **AI-powered regression model:** An online machine learning prediction model is built using advanced regression methods such as Bayesian regression, XGBoost regression, and others.
- **Reaction microkinetic modeling:** Enables real-time prediction of reaction kinetics based on experimental and theoretical datasets.
- **Machine-learning force field development:** Supports accurate simulations of catalytic processes under realistic conditions.

So far, some notable works have been published based on the new materials phenomenon identified by the DigCat. For example, based upon big data, the DigCat identifies the anomalously high oxygen reduction activity of weak-binding M–N–C single-atom catalysts [10] and the pH-dependent performance of Sn-based CO₂RR catalysts [11,12], which are almost brand-new insights in the area of electrocatalysis. Besides, the DigCat has been used for the comparative and benchmarking analysis by comparing the performance between the new materials and the literature materials (e.g., for catalytic water purification [13] and electrocatalytic hydrogen evolution [14] and oxygen evolution [15]). “Finding new insights from old papers” become possible.

3.2 DDSE: The Largest Dynamic Solid-State Electrolyte Database

The Dynamic Database of Solid-State Electrolytes (DDSE: <https://www.ddse-database.org/>) is a unique resource for solid-state battery research, containing over 2500 experimentally validated solid-state electrolytes and 600 computationally predicted candidates [5,6]. DDSE provides a foundation for

accelerating the development of next-generation battery materials by integrating experimental and AI-driven insights.

Key innovations of DDSE:

- **AI-driven conductivity prediction:** Utilizes machine learning models to identify promising SSEs for all-solid-state batteries.
- **User-interactive material comparison:** Enables researchers to input experimental data and compare against existing entries.
- **Integrated Large Language Model (LLM) analytics:** Automates literature analysis to extract performance trends and guide material selection.

A comparative overview of major materials databases, including DDSE and DigCat, is provided in [Table 1](#).

Table 1: Comparative overview of typical materials databases

Material databases	Types	Materials	License	AI functions
Inorganic Crystal Structure Database (ICSD)	Experimental	Inorganic crystal	Paid	None
Cambridge Structural Database (CSD)	Experimental	Organic molecules	Paid	None
Materials Project (MP)	Computational	Inorganic crystal	Free	AI regression model
Automatic-FLOW (AFLOW)	Computational	Inorganic crystal	Free	AI regression model
ElectroCat	Experimental	Fuel cell catalyst	Free	None
Catalysis-Hub.org	Computational	Thermo and electrocatalysts	Free	AI regression model
MatNavi	Experimental	Polymers, inorganics, metals	Free	None
AtomWork Adv	Experimental	Crystal structure	Paid	None
DigCat	Experiment + Computational	Catalysts	Free	AI Q&A, Machine Learning regression model, Machine learning force field, AI data collection, LLM-based Chatbot
DDSE	Experiment + Computational	Solid state electrolytes	Free	AI Q&A, Machine Learning regression model, Machine learning force field, AI data collection, LLM-based Chatbot

4 Toward the AlphaFold of Materials Science

DigCat and DDSE represent the next frontier in AI-driven materials discovery, analogous to AlphaFold's impact on protein folding prediction. However, the complexity of materials science far surpasses that of protein structures, particularly in areas such as:

- **Surface interactions and heterogeneous catalysis:** Unlike proteins, which operate in well-defined biological environments, catalytic materials function under diverse conditions, requiring sophisticated modeling of adsorption/desorption kinetics and reaction pathways.
- **Dynamic and nonequilibrium behaviors:** Battery electrolytes, for instance, exhibit non-static behaviors influenced by external fields, making accurate AI predictions significantly more challenging.
- **Interfacial phenomena:** The performance of solid-state batteries and heterogeneous catalysts depends on **interface stability, defect dynamics, and long-term degradation mechanisms**, which are difficult to capture using static datasets.

By leveraging vast datasets and advanced AI methodologies, DigCat and DDSE are pioneering the first truly predictive materials discovery platforms—a step toward achieving an “AlphaFold for materials.” Unlike AlphaFold, which operates within a well-defined sequence-structure relationship, materials databases must contend with multi-variable dependencies, environmental effects, and synthesis constraints, making their predictive capabilities even more groundbreaking.

5 Conclusion

AI-driven materials databases are reshaping the research landscape, moving from passive data repositories to intelligent, self-updating platforms that accelerate discovery. DigCat and DDSE, as pioneering initiatives, exemplify this transformation, offering unprecedented predictive power and dynamic insights. With continued advancements in machine learning, automated synthesis, and high-performance computing, the vision of a truly autonomous materials discovery engine is becoming a reality—one that may surpass even the impact of AlphaFold by tackling the unparalleled complexity of materials science.

Acknowledgement: The authors acknowledge Prof. Qiang Wang (Chinese Academy of Sciences) for the helpful discussions for this paper.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Yaping Qi, Weijie Yang; data collection: Yaping Qi, Weijie Yang; analysis and interpretation of results: Yaping Qi, Weijie Yang; draft manuscript preparation: Yaping Qi, Weijie Yang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Jain A, Ong S, Hautier G, Chen W, Richards W, Dacek S, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* 2013;1(1):011002. doi:10.1063/1.4812323.
2. Belsky A, Hellenbrandt M, Karen VL, Luksch P. New developments in the inorganic crystal structure database (ICSD): accessibility in support of materials research and design. *Acta Crystallogr B.* 2002;58(3–1):364–9. doi:10.1107/S0108768102006948.

3. Curtarolo S, Setyawan W, Hart GLW, Jahnatek M, Chepulskii RV, Taylor RH, et al. AFLOW: an automatic framework for high-throughput materials discovery. *Comput Mater Sci.* 2012;58:218–26. doi:10.1016/j.commatsci.2012.02.005.
4. Zhang D, Li H. Digital catalysis platform (DigCat): a gateway to big data and ai-powered innovations in catalysis. *ChemRxiv.* 2024. doi: 10.26434/chemrxiv-2024-9lpb9.
5. Yang F, Campos dos Santos E, Jia X, Sato R, Kisu K, Hashimoto Y, et al. A dynamic database of solid-state electrolyte (DDSE) picturing all-solid-state batteries. *Nano Mater Sci.* 2024;6(2):256–62. doi:10.1016/j.nanoms.2023.08.002.
6. Yang F, Wang Q, Cheng EJ, Zhang D, Li H. User instructions for the dynamic database of solid-state electrolyte 2.0 (DDSE 2.0). *Comput Mater Contin.* 2024;81(3):3413–9. doi:10.32604/cmc.2024.060288.
7. Wang Z, Zheng YR, Chorkendorff I, Nørskov JK. Acid-stable oxides for oxygen electrocatalysis. *ACS Energy Lett.* 2020;5(9):2905–8. doi:10.1021/acsenerylett.0c01625.
8. Jia X, Li H. Data mining of stable, low-cost metal oxides as potential electrocatalysts. *Artif Intell Chem.* 2024;2(1):100065. doi:10.1016/j.aichem.2024.100065.
9. Jia X, Yu Z, Liu F, Liu H, Zhang D, Campos Dos Santos E, et al. Identifying stable electrocatalysts initialized by data mining: Sb₂WO₆ for oxygen reduction. *Adv Sci.* 2024;11(5):e2305630. doi:10.1002/advs.202305630.
10. Zhang D, She F, Chen J, Wei L, Li H. Why do weak-binding M-N-C single-atom catalysts possess anomalously high oxygen reduction activity? *J Am Chem Soc.* 2025;147(7):6076–86. doi:10.1021/jacs.4c16733.
11. Wang Y, Zhang D, Sun B, Jia X, Zhang L, Cheng H, et al. 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. doi:10.1002/anie.202418228.
12. Guo Z, Wang T, Liu H, Jia X, Zhang D, Wei L, et al. Electrochemical CO₂ reduction on SnO: insights into C₁ product dynamic distribution and reaction mechanisms. *ACS Catal.* 2025;15(4):3173–83. doi:10.1021/acscatal.4c07987.
13. Zhong KQ, Yu FY, Zhang D, Li ZH, Xie DH, Li TT, et al. Data-driven accelerated discovery coupled with precise synthesis of single-atom catalysts for robust and efficient water purification. *Angew Chem Int Ed.* 2025;e202500004. doi:10.1002/anie.202500004.
14. Zhou K, Liu H, Liu Z, Li X, Wang N, Wang M, et al. W-mediated electron accumulation in Ru–O–W motifs enables ultra-stable oxygen evolution reaction in acid. *Angew Chem Int Ed.* 2025;e202422707. doi:10.1002/anie.202422707.
15. Zhu A, Qiao L, Liu K, Gan G, Luan C, Lin D, et al. Rational design of precatalysts and controlled evolution of catalyst-electrolyte interface for efficient hydrogen production. *Nat Commun.* 2025;16(1):1880. doi:10.1038/s41467-025-57056-6.