PERSPECTIVE

National Science Open
4: 20250039, 2025
https://doi.org/10.1360/nso/20250039

Chemistry

Special Topic: Artificial Intelligence and Energy Revolution

Overcoming data scarcity challenges in AI-driven energy chemistry research

Yu-Hang Yuan[1], Yu-Chen Gao[1], Xiang Chen[1,2,*] & Qiang Zhang[1,2,3,*]

[1] *Beijing Key Laboratory of Complex Solid State Batteries, Department of Chemical Engineering, Tsinghua University, Beijing 100084, China;*
[2] *Innovation Center for Smart Solid-State Batteries, Yibin 644002, China;*
[3] *State Key Laboratory of Chemical Engineering and Low-Carbon Technology, Tsinghua University, Beijing 100084, China*

*Corresponding authors (emails: xiangchen@mail.tsinghua.edu.cn (Xiang Chen); zhang-qiang@mails.tsinghua.edu.cn (Qiang Zhang))

Received 8 August 2025; Revised 27 September 2025; Accepted 11 October 2025; Published online 18 November 2025

Artificial intelligence (AI) has become an increasingly important propellant for energy materials and energy chemistry research, such as accelerating advanced energy materials discovery [1], analyzing vast amounts of data from both experiments and computations [2], process optimization for materials syntheses, management and monitoring of energy storage devices such as lithium batteries, and algorithm-optimized grid load forecasting. Looking back at recent pioneering works of AI-driven energy chemistry research, constructing a dataset with both large quantity and high quality is almost the first step and largely determines the following success of training AI models and figuring out corresponding scientific issues.

Generally, data are usually produced by experiments or computations, such as density functional theory (DFT) calculations, molecular dynamics (MD) simulations, and experimental data collected by various scientific instruments. To date, there are more than 150 specialized chemical or materials datasets that have been constructed and documented in major repositories. For instance, NIST Chemistry WebBook provides thermochemical, thermophysical, and ion energetics data. The Open Catalyst Experiments 2024 (OCx24) dataset combines computational descriptors and experimental outcomes for heterogeneous catalysis. The CatMath platform enables the computational construction and visualization of Pourbaix diagrams and catalytic volcano models, covering several important reactions in both electrocatalysis and thermal catalysis [3]. Unfortunately, the current data in the energy chemistry field is extremely far from enough, especially compared with AI studies contributed from the field of computer science. More miserably, data crossing from different published datasets are often not comparable, which impedes a highly efficient utilization of existing scarce data.

To overcome the above data scarcity challenges in energy chemistry research, four solutions are discussed with both recent important progress and promising future development directions (Figure 1). Among them, high-throughput computation and experimentation are still domain approaches for constructing a large dataset. Simultaneously, emerging AI technologies afford new opportunities for building datasets by text mining and data augmentation.

High-throughput computation is still the most convenient approach to constructing a dataset, especially for materials datasets. For example, the Materials Project database [4], started from 2011, has obtained 200,487 materials and 577,813 molecule entries through high-throughput first-principles calculations (version 2025.06.09). AFLOW dataset [5] includes 734,308,640 calculated properties of 3,530,330 inorganic crystals. The OQMD dataset [6] contains thermodynamic and structural properties of 1,317,811 inorganic materials by DFT calculations. Chen, Zhang, and co-workers [7] have constructed a battery electrolyte dataset with more than 250 thousand molecules and more than 20 physicochemical properties for each molecule, including dielectric constant, viscosity, diffusivity, ionic conductivity, dipole moment, and molecular frontier orbital energy level. These databases afford programmatic access via API and/or embedded machine learning models, facilitating the application of AI in further computations. However, current datasets from similar computations are facing potential discrepancies from experiments or between each other, which is caused by various computation methods or empirical parameters. Therefore, the future development of computational datasets necessitates the implementation of standardized data generation, representation, and exchange protocols, including adopting consistent symbol representations, universally recognized units, and machine-readable data export formats.

Experiments are still indispensable for obtaining many physicochemical properties and almost all device performances. For instance, atomic simulations are facing challenges in predicting the experimentally obtainable flash points of liquids. Conventional experimental datasets, such as Reaxys and PubChem, were often constructed by repeated manual experiments and collections. High-throughput experimentation, especially integrated with chemical robots, affords emerging opportunities for constructing experimental datasets. For example, Jiang *et al*. [8] built an AI-Chemist to leverage high-throughput experimentation to rapidly generate a dataset comprising validated, iteratively optimized experiment workflows and corresponding experiment results containing 207 overpotential measurements for oxygen evolution reaction electrocatalysts. Szymanski *et al*. [9] built an A-Lab platform that optimizes synthesis routes through active learning and builds a database of pairwise reactions by continuously performing 355 experiments over 17 days. Cooper and coworkers [10] developed a mobile robotic chemist that obtains a multi-dimensional database by integrating real-time multimodal characterizations (nuclear magnetic resonance and mass spectrometry). Collectively, these examples underscore how high-throughput experimentation platforms not only expedite material discovery but also function as powerful database engines. These platforms are expected to systematically generate, compile, and utilize vast amounts of structured experimental data, with the inherent data quality further enhanced by cross-referencing with established databases and proactive equipment maintenance.

Given the exponentially growing body of published papers and patents containing experimental and/or computational data, text mining presents an alternative approach to construct a dataset using existing data rather than producing new data. For example, in 2024, Web of Science indexed approximately 2.5 million articles in the field of chemistry, materials science, energy fuels, and electrochemistry. To read, comprehend, and summarize the numerous articles, large language models (e.g., ChatGPT and DeepSeek) serve as an efficient tool with extensive knowledge and powerful key information extraction capability, enhancing both the reliability and scalability of text mining. Leveraging AI-driven pipelines, text mining can efficiently distill and synthesize knowledge from this vast body of literature. The typical workflow involves [11] content

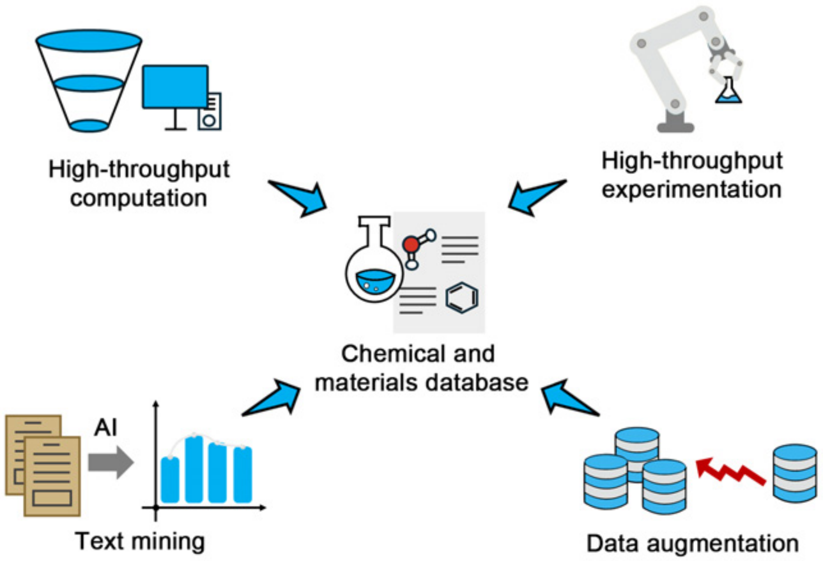
High-throughput computation
High-throughput experimentation
Chemical and materials database
AI
Text mining
Data augmentation

Figure 1 Schematic diagram of four strategies to efficiently obtain an energy chemistry database.

acquisition and markup parsing, text pre-processing, document segmentation, entity recognition, and knowledge graph building (optional). For instance, Zheng *et al*. [12] proposed ChatGPT Chemistry Assistant, which extracted 26,257 synthesis parameters from approximately 800 research articles about metal-organic frameworks (MOFs). Utilizing these mined data, a machine learning model is constructed to predict MOF experimental crystallization outcomes with an accuracy of over 87%. However, consistency and repeatability remain major challenges in reproducing experimental data from papers. To address these issues, it is essential to document the specific testing conditions under which experimental data are obtained. Collecting literature that reports data under similar conditions facilitates meaningful comparisons, while incorporating data obtained under extreme or unique conditions enhances the scalability and breadth of the database.

Data augmentation remains a last potential choice for the AI study of energy chemistry when no more data is available from the above computations, experiments, or published sources. New synthetic data can be generated by domain-appropriate transformations, of which the relationship between inputs and targets can be learned by data augmentation models trained on existing data. Both labeled data and unlabeled data can be generated by data augmentation methods, such as mixup interpolation and pseudo-labeling, respectively. For instance, Jiang *et al*. [13] augmented the training data for AI models by adopting proper token operations (mask, swap, deletion, and fusion) on the SMILES (a widely used molecular representation method) and enhanced the prediction capability of AI models. In microscopic imaging, data augmentation is achieved through the fusion of real and simulated images. A deep learning model trained on the derived hybrid dataset (35% real image and 65% synthetic image) demonstrates competitive performance compared to its counterpart trained exclusively on actual data, showcasing the significant potential of data augmentation [14]. Despite enabling the rapid establishment of large-scale databases, it is hard to estimate the quality of synthetic data, which remains a major challenge for data augmentation.

In summary, the synergistic integration of the above four strategies affords a comprehensive solution to address the data scarcity challenges in AI-driven energy chemistry research. High-throughput computation and experimentation fundamentally generate new data, while text mining and data augmentation afford powerful tools for taking full advantage of existing valuable data. Data generated by high-throughput experimentation can inform computational workflows in high-throughput computation through methods such as Bayesian optimization, reinforcement learning, and active learning [15]. To identify and manage erroneous or low-quality data, it is essential to apply data visualization and evaluation techniques. Visualization methods (e.g., t-SNE, PCA) and data analysis approaches (e.g., Shapley analysis) can assist researchers in detecting anomalies and assessing data quality by uncovering structural outliers in data projections and quantifying the contribution of individual factors to model predictions, respectively. For subsequent data cleaning, tools such as ActiveClean can be employed to enhance efficiency. Looking forward, enhanced sharing, analysis, and optimization across all four strategies, along with new data generation and processing systems (e.g., integrated human-AI data systems), is anticipated to mitigate the challenge of data scarcity in energy chemistry research.

Funding

This work was supported by the National Key Research and Development Program of China (2021YFB2500300), the National Natural Science Foundation of China (T2322015, 92472101, 22393903, 22393900, 52394170), and the Beijing Municipal Natural Science Foundation (L247015, L233004).

Conflict of interest

The authors declare no conflict of interest.

References

1 Gao Y, Yuan Y, Huang S, *et al.* A knowledge-data dual-driven framework for predicting the molecular properties of rechargeable battery electrolytes. *Angew Chem Int Ed* 2025; **64**: e202416506.

2 You Q, Sun Y, Wang F, *et al.* Decoding the competing effects of dynamic solvation structures on nuclear magnetic resonance chemical shifts of battery electrolytes via machine learning. *J Am Chem Soc* 2025; **147**: 14667–14676.

3 Liu H, Zheng H, Jia Z, *et al.* The CatMath: An online predictive platform for thermal + electrocatalysis. *Front Chem Sci Eng* 2023; **17**: 2156–2160.

4 Jain A, Ong SP, Hautier G, *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater* 2013; **1**: 011002.

5 Calderon CE, Plata JJ, Toher C, *et al.* The AFLOW standard for high-throughput materials science calculations. *Comput Mater Sci* 2015; **108**: 233–238.

6 Kirklin S, Saal JE, Meredig B, *et al.* The open quantum materials database (OQMD): Assessing the accuracy of DFT formation energies. *npj Comput Mater* 2015; **1**: 15010.

7 Chen X, Liu M, Yin S, *et al.* Uni-electrolyte: An artificial intelligence platform for designing electrolyte molecules for rechargeable batteries. *Angew Chem Int Ed* 2025; **64**: e202503105.

8 Zhu Q, Zhang F, Huang Y, *et al.* An all-round AI-Chemist with a scientific mind. *Natl Sci Rev* 2022; **9**: nwac190.

9 Szymanski NJ, Rendy B, Fei Y, *et al.* An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* 2023; **624**: 86–91.

10 Dai T, Vijayakrishnan S, Szczypiński FT, *et al.* Autonomous mobile robots for exploratory synthetic chemistry. *Nature* 2024; **635**: 890–897.

11 Olivetti EA, Cole JM, Kim E, *et al.* Data-driven materials research enabled by natural language processing and

information extraction. *Appl Phys Rev* 2020; **7**: 041317.

12 Zheng Z, Zhang O, Borgs C, *et al.* ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *J Am Chem Soc* 2023; **145**: 18048–18062.

13 Jiang J, Zhang R, Yuan Y, *et al.* NoiseMol: A noise-robusted data augmentation via perturbing noise for molecular property prediction. *J Mol Graphics Model* 2023; **121**: 108454.

14 Ma B, Wei X, Liu C, *et al.* Data augmentation in microscopic images for material data mining. *npj Comput Mater* 2020; **6**: 125.

15 Wang JY, Stevens JM, Kariofillis SK, *et al.* Identifying general reaction conditions by bandit optimization. *Nature* 2024; **626**: 1025–1033.