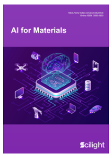
AI for Materials
https://www.sciltp.com/journals/aimat



Review

Let Materials Science Data Learn to Reason

Tongao Yao [1,2], Aoni Xu [3,4], Pengfei Ou [5,6] and Weijie Yang [1,2,]*

[1] Department of Power Engineering, North China Electric Power University, Baoding 071003, China
[2] Hebei Key Laboratory of Energy Storage Technology and Integrated Energy Utilization, North China Electric Power University, Baoding 071003, China
[3] School of Chemical and Biomolecular Engineering, The University of Sydney, Sydney, NSW 2006, Australia
[4] ARC Centre of Excellence for Green Electrochemical Transformation of Carbon Dioxide, The University of Sydney, Sydney, NSW 2006, Australia
[5] Department of Chemistry, National University of Singapore, 3 Science Drive 3, Singapore 117543, Singapore
[6] NUS Artificial Intelligence Institute, National University of Singapore, 11 Research Link, Singapore 119391, Singapore
* Correspondence: yangwj@ncepu.edu.cn

How To Cite: Yao, T.; Xu, A.; Ou, P.; et al. Let Materials Science Data Learn to Reason. *AI for Materials* **2026**, *1*(1), 4.

Received: 6 October 2025
Revised: 7 December 2025
Accepted: 17 December 2025
Published: 22 December 2025

Abstract: Materials science excels at collecting numbers but often loses sight of the conditions, mechanisms, and provenance that give those numbers meaning. We have vast databases, long papers, and a forest of plots, yet the bench-top questions persist: what works here, now, under these conditions and why? This Viewpoint argues for a practical posture: treat conditions as first principles, make mechanisms explicit, and attach sources and uncertainty to every claim. We outline a plain work loop, i.e., ask → retrieve → compute → validate → write back, that increases learning speed without requiring "hero" datasets or black-box models. We demonstrate how these habits are applied in practice across catalysis, solid-state batteries, and hydrogen storage, and we highlight platform design choices (e.g., DigMat Platform) that enforce them. The goal is not fireworks, but fewer wrong turns: fairer comparisons, predictions tied to the exact world they apply to, and decisions that withstand daylight.

Keywords: *AI for Materials*; condition-first reasoning; data provenance; mechanism-constrained inference; out-of-distribution

1. From Piling up Numbers to Asking Better Questions

We long treated data like "ore" [1]: we dug it up, stacked it, and hoped value would appear by sheer volume. It rarely does. Real lab questions are local and conditional: *Why does this surface wake up only at high pH? Why does a half-cell result fade inside a real device? If we can change only two knobs before lunch, which two should they be?* Those questions never felt "big", so we assumed the answers would be easy. They are not. Numbers without context rarely provide clear guidance on what to try next.

Vague questions—"Does catalyst A beat B?"—invite confident but unhelpful answers. Precise questions carry the context that makes answers actionable: "At pH 13, with this electrolyte, under this transport regime, does catalyst A outperform B at a given overpotential?". Concurrently, when precise questions meet emerging techniques, LLM-enabled retrieval/reasoning [2] and generative design frameworks [3], the benefits become apparent when embedded in a condition-first, provenance-aware practice. Precision is not pedantry; it is how we avoid comparing apples to something that merely looks like an apple.

While the community has rightly embraced FAIR principles (Findable, Accessible, Interoperable, Reusable) to organize data, we argue that accessibility alone is insufficient for reasoning. FAIR ensures a number can be found; it does not ensure the number is physically valid for a specific inference context. Standard databases often treat data as static records. In contrast, we propose a dynamic paradigm termed "Condition-First Reasoning", where physical mechanisms and environmental constraints are hard-coded into the retrieval and inference logic.

This shift is critical to prevent AI models from hallucinating in “Out-of-Distribution” (OOD) regions, a challenge that metadata tagging alone cannot solve.

2. Conditions Are Not Footnotes

Many repositories record what we tested, but often forget the specific conditions under which it was done. That is how “apples-to-apples” usually becomes “apples-to-something-that-looks-like-an-apple”. Electrolyte, potential or current mode, pH, film thickness, surface state, stirring or flow, temperature, pressure: these are not afterthoughts buried in the supplement. They are the stage on which the play happens. If you move the stage, the same script gives a different story.

Imagine a student reading two papers that both claim “excellent stability”. One ran at 25 °C in a calm beaker; the other at 60 °C under flow. The charts look similar; the realities are not. When a database treats conditions as first-class, the student no longer needs detective work. They can filter by the world they care about, not the one a figure silently assumed.

In this Viewpoint, “treating conditions as first principles” means that environmental variables such as pH, temperature, pressure, applied potential, mass-transport regime, and surface state are treated as primary inputs that define the validity window of any descriptor, mechanism, or prediction. These conditions are embedded directly in the query structure and modeling feature space, constraining inference in the same way that boundary conditions constrain a differential equation. A prediction is therefore considered valid only when its provenance vector matches the query condition vector. This practice prevents out-of-distribution extrapolation and ensures that comparisons remain tied to the physical world in which they are meant to hold.

3. Make Mechanism Visible

Mechanism turns lists into stories. In practice, we reason through a web of descriptors and limits, including adsorption, coverage, barriers, rate-determining steps, phase stability, defects, and mass transport limits. Each one matters only within a window. If we hide those links, it is easy to sound convincing for the wrong reason.

Let us consider the oxygen reduction reaction (ORR) [4]. A simple descriptor might explain a trend at one pH and fail at another when surface coverage flips or mass transfer takes over. When the chain is explicit; *this holds at high pH because this intermediate dominates and the transport cap moves here*; we can argue less and test faster. A claim you cannot trace is not a claim; it is a hunch.

4. Multi-Scale Integration

Mechanisms extend beyond the microscale: they must be visible across scales. From atomic adsorption to macroscopic degradation, each layer limits couple tightly. AI platforms should build in multi-scale mappings, like linking DFT barriers to battery cycle fade, otherwise paper mechanisms never become real experiments.

5. Provenance, Not Performance

Science is full of confident sentences. Confidence is cheap; provenance is expensive and worth it. Every answer should carry its sources, assumptions, and error bars. “Most promising” means little unless it also says “according to these papers, under these windows, with these knobs, within this uncertainty”. This is not red tape; it is a time saver. Provenance turns debates into choices: *if we disagree, which assumption should we flip first? What measurement would settle it by Friday?* Provenance reduces the drama and amplifies the signal-to-noise ratio.

6. A Plain, Workable Loop

Once we accept that context and mechanism matter, the workflow becomes simpler and calmer.

We begin with a compute-ready question that has explicit conditions. We pull only evidence that matches the conditions. We run the calculation that the mechanism asks for, not the one that happens to be convenient. We test on a clear system, something robust and boring, before chasing complexity. Then we write the results back, including failures, so tomorrow search begins wiser than today.

Nothing here is glamorous, and that is the point. The loop is designed to reduce waste. It does not promise breakthroughs on demand; it promises fewer wrong turns. The full workflow is summarized visually in Figure 1, which illustrates how questions, evidence, computation, validation, and “write-back” interact in a closed feedback system to progressively reduce uncertainty.

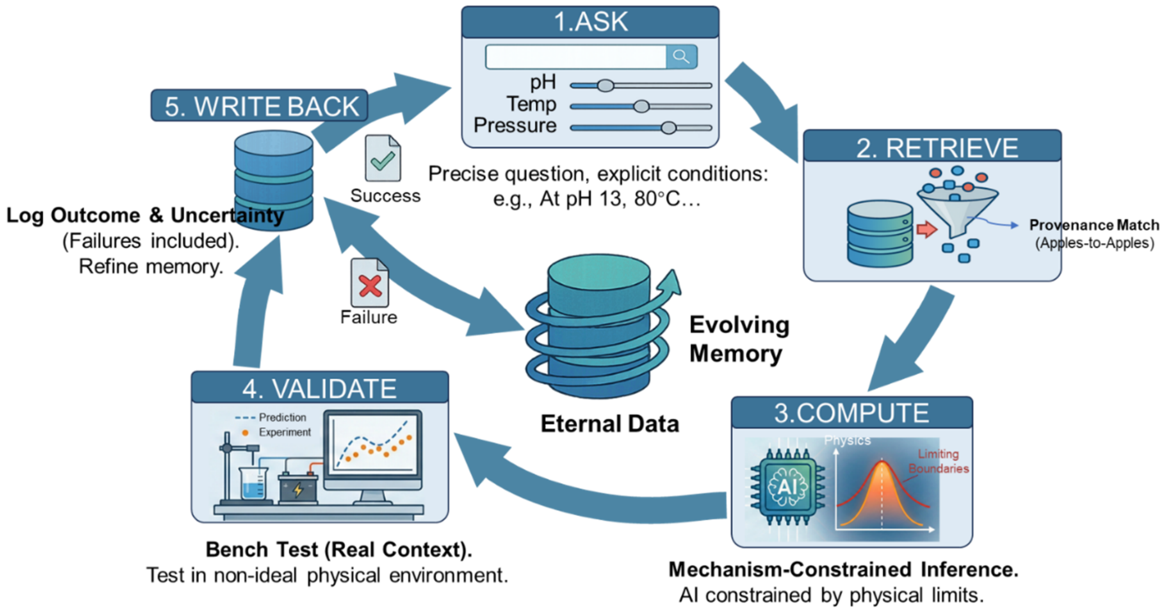


Figure 1. The Reasoning Loop for Materials Data. Instead of linearly piling up data, the workflow follows a closed loop: (1) Ask: Questions define explicit environmental conditions (e.g., pH, T) upfront. (2) Retrieve: Only provenance-matched evidence is pulled; incompatible data is filtered out. (3) Compute: AI models are constrained by physical mechanisms (e.g., thermodynamic limits) rather than pure statistical extrapolation. (4) Validate: Predictions are tested in the exact physical context defined in step 1. (5) Write Back: Both successes and failures are logged with their specific conditions and uncertainties, refining the system's memory for the next cycle.

7. Learning Honestly in Messy Systems

Real materials behave like real life: coupled, noisy, and full of surprises. Interfaces are not ideal; manufacturing introduces constraints that no chart predicted; degradation appears after the tenth cycle and then persists. In this setting, the aim should be modest and practical: compare fairly, predict in context, and decide with costs and risks accounted for.

When we connect this discipline to routine automation, such as simple synthesis, standard diagnostics, and quick-turn tests, it becomes a bench-side companion. Not a blinking dashboard, not a "black box", just a system that remembers, checks, and nudges.

8. The OOD Trap in AI-Empowered Materials Design

AI-empowered materials design almost always faces a common pitfall: the uncertainty of out-of-distribution (OOD) predictions. Existing databases are filled with known materials, those mediocre, even "poor" performers, that form the model training distribution. But what we truly seek is the realm of excellence: the "blank zone" of high capacity, low cost, and ultra-stability, where data points are scarce. When models extrapolate to these unknown regions, biases naturally swell, a seemingly "optimal" prediction may just be a blind extension of low-performance trends, leading to wasted experiments and false positives.

This is not the "fault" of AI, but a reflection of data nature: history favors the readily available losers over the rare winners. The remedy lies in honestly acknowledging OOD: attach distribution confidence intervals to predictions, and prioritize physical mechanisms to constrain models (like thermodynamic limits) over pure data fitting. At the same time, introduce active learning loops, letting the system generate high-value experiment suggestions targeted at OOD regions, and write back after validation, so the blanks gradually fill, and predictions grow reliable. In the end, this strengthens the original loop: not just asking "what is optimal", but "under OOD, how inaccurate is this optimal, and why?".

9. Platforms as Habits: The DigMat Example

Platforms matter only if they help us keep these habits. The Digital Materials Platform (DigMat: www.digmat.org (accessed on 18 December 2025)), developed by Hao Li and colleagues at Tohoku University, is one representative example. Its design choices are straightforward: place conditions alongside composition, not beneath it; maintain a living map of mechanisms and descriptors; ensure every answer includes sources and uncertainty; and log every outcome back into the system. To support this condition-first posture, the platform organizes its data in a structured schema rather than a flat table. Each record links a material system to a condition vector that specifies parameters such as pH, temperature, pressure, applied potential, transport regime, and surface state. Properties are associated with mechanism labels that indicate the relevant operating regime, and every entry carries explicit provenance including source identifiers and uncertainty.

The DigMat platform then applies the same rhythm in three areas. In DigCat (catalysis: www.digcat.org (accessed on 18 December 2025)) [5–7], pH, potential, and transport are built into the way we ask and answer, so predictions are tied to microkinetic reasoning before an experiment starts. In DigBat (solid-state batteries: www.digbat.org (accessed on 18 December 2025)) [8,9], ion transport and interfacial compatibility coexist with conductivity, which prevents "paper wins" that fail upon contact. In DigHyd (hydrogen storage: www.dighyd.org (accessed on 18 December 2025)) [10], the ranking balances capacity, kinetics, cycling, and cost as actually tested, not as marketed. Pilot-scale results are allowed to correct earlier optimism. None of these tries to crown a winner. It tries to keep our bets honest.

This condition-first ethos shines in hydrogen storage platforms like XPEAK [11] and Cat-Advisor [12] for catalysis, FIND [13] for alloying, all anchored in the open Digital Hydrogen-S platform [14] (digital-hydrogen.com/storage (accessed on 18 December 2025)). The database curates preparation, characterization, and performance data across Mg-based, AB, and related systems, enabling automated literature benchmarking for fair comparisons. Crucially, these implementations demonstrate that enforcing logic yields performance gains that standard models cannot match. XPEAK uses more than 2000 MgH_2 XRD fingerprints to predict dehydrogenation peaks ($R^2 = 0.86$); this precision was achieved by strictly anchoring predictions to valid experimental domains, whereas unconstrained models often fail to distinguish signal from noise. Similarly, Cat-Advisor curates 809 catalysts via LLM ($R^2 > 0.91$), using genetic algorithms not just to optimize, but to align predictions with physical validity windows. FIND integrates over 6000 records with variational autoencoders to optimize plateau pressures, enthalpies, and capacities in Mg-Ni-La-Ce systems. Together, these tools enforce precise questions, explicit mechanisms, and traceable provenance, yielding contextual predictions and fewer wrong turns.

10. Keep Intuition—Add Memory

Good groups already behave this way in their heads. They carry context, prefer mechanisms over buzzwords, distrust claims without lineage, and close loops fast. A shared system adds what people cannot: durable memory. It preserves the awkward footnotes that would otherwise get lost; records why a nice alloy failed when the electrolyte changed; saves negative results that never made a figure but would have saved a month; and lets a new student inherit judgment, not just folders.

11. Human-AI Symbiosis

Memory augments intuition but does not replace it. AI excels at patterns; humans at "why not". The ideal collaboration: let AI propose hypotheses, humans inject domain knowledge (like tacit manufacturing constraints), and together scrutinize OOD risks. Thus, the system is no oracle, but a sparring partner, helping us iterate faster.

12. Two Small Stories

ORR screen with context. A descriptor-only model applied to Pt_3Ni returns a top-five list for ORR activity (0.9 V overpotential); two fail in alkaline electrolyzer cells (pH 13, 60 °C flow) due to unaccounted HO* coverage shifts promoting a peroxide pathway [15]. Adding microkinetic mechanisms reranks the list to favor HO*-tolerant alloys like Pt_3Co; now four out of five achieve >80% stability over 100 h under the same regime. This shift proves the viewpoint core: pH-transport context yields provenance-aware predictions, turning hunches into validated wins.

A "perfect" solid electrolyte. $Li_7La_3Zr_2O_{12}$ (LLZO) performs well in room-temperature half-cells (25 °C, 10–14 S/cm conductivity, 100 cycles < 5% fade); yet in full stacks (80 °C, 1C rate), interfacial voids cause failure after 20 cycles [16]. A platform prioritizing compatibility/processing alongside conductivity flags this risk pre-build via DFT simulations. Not magic: condition-first retrieval and mechanism visibility ensure fair comparisons across half- to full-cell realities and avoid "paper-to-device" pitfalls.

13. Lighter Questions, Heavier Answers

We often debate about the "right" algorithm or the "best" dataset. Those arguments cool down when we agree to ask lighter questions with heavier answers. *What holds, when, and why?* If we can also say *how we would try to break the claim tomorrow*, progress starts to feel real. The field gets quieter in a good way.

14. Conclusions

Data-driven materials research should not mean “more data”. It should mean “better memory and better manners”. Do not lose context. Do not imply mechanism; show it. Do not make anonymous claims; attach a source and an uncertainty. When we work this way, we trade volume for traction: comparisons become fair; predictions stay tied to the world they belong to; Decisions stand up in daylight. For instance, the DigMat platform is effective because it encourages these habits without drawing attention to them. It takes a general posture: ask carefully, compute honestly, validate early, write back always; and lets it play out in three different families of problems. That is how a digital idea turns into ordinary practice: fewer grand promises, fewer wasted afternoons, more steady wins.

15. Outlook

There are a few things we can do right away. We can agree on standard reporting windows: shared, unglamorous ranges of pH, temperature, flow, and current where results must be given. This single step would shrink many cross-paper debates. We can also commit to writing back failures. Negative and unclear outcomes are cheap lessons that a field often hides; sharing them openly saves everyone time. Besides, we can tie modest automation to precise questions: rigs that test this claim under these conditions by Friday and log the outcome with its uncertainty.

Looking further ahead, the most interesting aspect is the transfer without distortion. Methods that behave in catalysis should be applied to batteries and hydrogen storage, and return with scars but still functioning. Three branches of DigMat are one conversation about that transfer. Other groups will build different shapes and names. If they keep the same habits: context first, mechanism explicit, provenance mandatory, loop closed, the labels will matter less. What will matter is that the bench grows steadily calmer, choices grow more grounded, and the next student starts one rung higher than the last.

Author Contributions

T.Y.: Conceptualization, methodology, formal analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization; A.X.: Methodology, validation, investigation, writing—review and editing; P.O.: Methodology, validation, investigation, writing—review and editing; W.Y.: Conceptualization, resources, writing—original draft preparation, writing—review and editing, supervision, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Natural Science Foundation of Hebei, grant number E2023502006, the Fundamental Research Funds for the Central Universities-China, grant number 2025JC008, and the Fundamental Research Funds for the Central Universities, grant number 2025MS131.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

No new data were generated in this article.

Conflicts of Interest

Given the role as editorial board member of *AI for Materials*, Pengfei Ou had no involvement in the peer review of this paper and had no access to information regarding its peer-review process. Full responsibility for the editorial process of this paper was delegated to another editor of the journal.

Use of AI and AI-Assisted Technologies

Use of AI and AI-Assisted Technologies During the preparation of this work, the authors used ChatGPT to improve readability and refine English language usage. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Qi, Y.P.; Yang, W.J. From Data to Discovery: How AI-Driven Materials Databases Are Reshaping Research. *Cmc-Comput. Mater. Contin.* **2025**, *83*, 1555–1559. https://doi.org/10.32604/cmc.2025.064061.
2. Yao, Y.W.; Zhu, J.B.; Liu, Y.; et al. Large Language Models for Heterogeneous Catalysis. *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **2025**, *15*, e70046. https://doi.org/10.1002/wcms.70046.
3. Chao, Y.; Lulu, W.; Jinbo, Z.; et al. Heterogeneous catalyst design by generative models. *J. Mater. Inform.* **2025**, *5*, 46.
4. Kulkarni, A.; Siahrostami, S.; Patel, A.; et al. Understanding Catalytic Activity Trends in the Oxygen Reduction Reaction. *Chem. Rev.* **2018**, *118*, 2302–2312. https://doi.org/10.1021/acs.chemrev.7b00488.
5. Zhang, D.; Li, H. The hidden engine of AI in electrocatalysis: Databases and knowledge graphs at work. *Mol. Chem. Eng.* **2025**, *1*, 100003. https://doi.org/10.1016/j.mochem.2025.100003.
6. Di, Z.; Xue, J.; Heng, L.; et al. Cloud synthesis: A global close-loop feedback powered by autonomous AI-driven catalyst design agent. *AI Agent* **2025**, *1*, 2.
7. Zhang, D.; Li, H. Digital Catalysis Platform (DigCat): A Gateway to Big Data and AI-Powered Innovations in Catalysis. *ChemRxiv* **2024**. https://doi.org/10.26434/chemrxiv-2024-9lpb9.
8. Yang, F.L.; Wang, Q.; Cheng, E.J.; et al. User Instructions for the Dynamic Database of Solid-State Electrolyte 2.0 (DDSE 2.0). *Cmc-Comput. Mater. Contin.* **2024**, *81*, 3413–3419. https://doi.org/10.32604/cmc.2024.060288.
9. Wang, Q.; Yang, F.L.; Wang, Y.H.; 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*, e202506573. https://doi.org/10.1002/anie.202506573.
10. Zhang, D.; Jia, X.; Hung, T.B.; et al. “DIVE” into Hydrogen Storage Materials Discovery with AI Agents. *arXiv* **2025**, arXiv:2508.13251.
11. XPEAK Platform. An XRD-Driven Machine Learning Tool for Predicting the Dehydrogenation Peak Temperature of MgH_2. Available online: cat-mh.top (accessed on 19 December 2025).
12. Yao, T.; Yang, Y.; Cai, J.; et al. From LLM to Agent: A Large-Language-Model-Driven Machine Learning Framework for Catalyst Design of MgH_2 Dehydrogenation. *J. Magnes. Alloys* **2025**, in press. https://doi.org/10.1016/j.jma.2025.08.021.
13. Xuao, L.; Shiwen, L.; Jiongyang, L.; et al. FIND: A forward–inverse navigation and discovery platform for hydrogen storage alloys powered by data-driven machine learning. *J. Surveill. Secur. Saf.* **2025**, *5*, 48.
14. Digital Hydrogen-S: An Open-Access Interactive Data Platform for Solid-State Hydrogen Storage Materials. Available online: digital-hydrogen.com/storage (accessed on 19 December 2025).
15. Luo, M.C.; Koper, M.T.M. A kinetic descriptor for the electrolyte effect on the oxygen reduction kinetics on Pt(111). *Nat. Catal.* **2022**, *5*, 615–623. https://doi.org/10.1038/s41929-022-00810-6.
16. Lou, S.F.; Zhang, F.; Fu, C.K.; et al. Interface Issues and Challenges in All-Solid-State Batteries: Lithium, Sodium, and Beyond. *Adv. Mater.* **2021**, *33*, 2000721. https://doi.org/10.1002/adma.202000721.