https://doi.org/10.1038/s41570-025-00750-2

Real AI advances require collaboration

N. M. Anoop Krishnan & Kevin Maik Jablonka

Standardized tasks and data sets have enabled machine learning experts to contribute to chemistry. However, current practices tend to reward incremental improvements in benchmark performance that do not necessarily translate into meaningful advances. We propose a framework for collaboration between artificial intelligence and domain experts to genuinely accelerate discovery.

Scientific progress depends on the sharing of knowledge. Journals share findings, conferences enable discourse and standardized methods ensure reproducible results. Such innovations have democratized scientific knowledge. Machine learning (ML) brought new standardization tools: benchmark data sets, evaluation metrics and leaderboards tracking incremental improvements. Progress in computer vision and natural language processing was driven by the creation of clear targets for improvement. Standardization meant that artificial intelligence (AI) researchers did not need deep domain expertise to contribute to science. This is not much different from how well-packaged tools (for example, for density functional theory or finite element simulations) made it easier for domain experts to perform such simulations themselves.

Yet this democratization came with unintended consequences. While AI researchers could engage more easily, domain experts found themselves excluded from the process[1,2]. The result is two parallel communities with insufficient dialogue: AI researchers work to optimize benchmark performance while domain experts question the relevance of those benchmarks for their real-world problems. There was no ill intent here — those closest to a problem seem best positioned to judge progress, but involving only experts can create echo chambers. Such a division has profound consequences if we wish to face challenges that demand integrated cross-discipline approaches, such as sustainability, the demand for clean energy and drug discovery. As the use of AI in science grows, we must examine what kind of communities enable discovery and whether current practices help or hinder their formation.

This Comment examines three problems: ML-ready training data sets that fail to capture real-world complexity, model evaluation frameworks disconnected from scientific relevance and institutional barriers preventing meaningful collaboration. Our proposed solutions focus on community integration: collaborative structures that value both AI and domain expertise and platforms that enable knowledge sharing combined with evaluation based on scientific impact rather than statistical performance.

The accessibility–progress paradox

Despite spectacular advances in general AI capabilities, applications in the chemical sciences and related fields have reached a troubling standstill[3,4]. We thus face a paradox: the increasing use of AI tools has not yielded proportional scientific breakthroughs. At the same time, the standardization that made these fields accessible to AI researchers has created barriers to meaningful scientific advancement. An analogy could be drawn to claims from computational scientists that experimental scientists "misuse" well-packaged simulation tools.

In the field of computer vision — a subfield of AI in which models interpret images — ImageNet (a large-scale database of approximately 14 million annotated images of, for instance, different vehicles or breeds of dog) catalysed genuine progress. But the complexity of scientific problems often defies such simplification. Benchmarks rely on comparison to verified correct answers, and real-world problems are rarely clear-cut. As a result, benchmark improvements seldom translate to discoveries. Consider materials property prediction or molecular force field development. Articles regularly claim improvements of several percentage points on standard data sets. Although these incremental advances suggest progress, they often lack the deeper insights needed for genuine scientific breakthroughs — chemical understanding, synthesizable materials with transformative properties, or new catalyst design[3,5].

The issue is not technical capability but community structure. Technical optimization is rewarded over scientific relevance, creating only an illusion of progress. The focus on creating 'well-packaged' forms of simplified scientific problems for AI has diverted the attention from the work on data sets and modelling systems that captures the real-world complexity.

Evaluation crisis

This illusion of progress stems from flawed evaluation frameworks. A telling example is provided by molecular force field development. Mean absolute error between predictions and ground truth energy or forces has become the default metric for judging success in these systems[6]. Yet this statistical measure often bears little relations to chemical relevance or practical utility. This manifestation of Goodhart's law — when a metric becomes a target, it ceases to be a good metric — has profound implications for chemical AI. Modern ML potentials report impressive accuracy on energy and force predictions for benchmark systems but fail catastrophically when applied to chemical configurations outside the training set[6]. The numbers improve while chemical understanding stagnates. Another case emerges in the use of AI to predict molecular trajectories or solutions to partial differential equations[3]. Articles report impressive accuracy in predicting individual molecular trajectories or the evolution of turbulent systems. Yet domain experts recognize that the ergodic nature of these systems means the precision of any single trajectory has limited value. Moreover, these models show poor transferability to new chemical systems. What matters are ensemble properties and whether predictions preserve fundamental physical and chemical principles — conservation of mass, momentum and energy; realistic thermodynamic behaviour; and correct reaction mechanisms. These chemical constraints rarely appear in benchmarks. Similar problems plague other areas, such as generative modelling of drug-like molecules[7] and catalyst discovery[2].

What has been unintentionally created is a benchmarking lottery in which problems with easily measurable outcomes attract disproportionate attention while the highest-leverage chemical activities — such as hypothesis generation and synthesizability — resist simple quantification. Our focus on easily measurable results disconnects us from discoveries that might transform chemical synthesis or materials design.

The chemical data crisis

Chemical knowledge lives in data, but many remain inaccessible, fragmented or unusable. Despite growing journal mandates for data sharing, compliance is low. Shared data often sit behind paywalls, locked in proprietary formats or stripped of the experimental context needed for reuse[8].

The data crisis in chemical AI runs deeper than technical accessibility. The most valuable chemical data — failed synthesis attempts that define feasibility boundaries, reaction conditions determining selectivity and stability measurements predicting shelf life — are created within experimental labs but remain invisible to AI researchers developing predictive models. This creates a vicious cycle: models trained on algorithmically convenient rather than chemically meaningful data produce results that chemists find irrelevant, further reducing their incentive to share their data. Without community integration, even perfect technical solutions for data sharing will fail to bridge this divide.

The burden of making chemical data reusable — curating synthesis protocols, documenting reaction conditions and aligning measurements with chemical principles — falls largely on chemists. Yet this work remains invisible in academic metrics. Our system rewards algorithmic novelty, not the infrastructure making chemistry cumulative. The problem transcends findable, accessible, interoperable and reusable (FAIR) data principles — it reflects fractured community structures. The forces isolating critical chemical data sets also exclude chemical expertise from AI development. When chemists cannot access or shape the data driving model design, their knowledge disappears from the loop. This disconnect weakens chemical grounding, undermines reproducibility and limits the collaborations we need most.

Fixing this requires more than better metadata. We must restructure how we value essential chemical contributions that do not result in leaderboard wins. Drawing from economic frameworks[9], we can envision scientific marketplaces recognizing the full spectrum of contributions, from chemical data curation to mechanistic interpretation.

Community engagement is the path forward

The future of AI for chemistry depends not on further democratization but on creating structures enabling true collaboration. We must unite AI researchers and chemists around problems of genuine chemical significance — moving beyond leaderboards into laboratories and pilot plants. We propose three directions:

Co-designed research environments. Collaborative frameworks need to value equally technical innovation, chemical insight, industry partnerships and stakeholder needs. We must create interdisciplinary labs with shared leadership and cross-disciplinary training so that domain experts learn how AI experts think and vice versa. Discipline-centred departments need to be transformed into problem-centred communities focused on challenges such as sustainable catalysis, battery materials or drug design.

Impact-centred evaluation. Realign evaluation metrics with broader, inclusive communities beyond developers — including chemists and downstream users[10]. Funders and publishers should prioritize challenge problems (for example, the Critical Assessment of Structure Prediction), co-defined with stakeholders from screen over bench to plant, and demand evidence of real-world utility rather than statistical performance alone.

Community infrastructure with shared stewardship. Build organizations and microgrant programmes that facilitate not just data sharing but active interdisciplinary exchange. Novel organizations could coordinate data generation: researchers submit proposals to generate chemical data sets and the data produced enter the public domain. Further, they could enable access to high-throughput experimentation, in a similar fashion to how access to beam time is currently managed. They could also organize competition-based evaluations for chemical problems. These could be facilitated by online platforms, which make it easy for domain experts to contribute new challenges or benchmarks.

These changes may face institutional resistance — departments questioning how to evaluate interdisciplinary work or researchers concerned about career risks. Overcoming these collective action problems requires coordinated effort: leadership from funding agencies and journals creating protected spaces for community integration, alongside grassroots demonstrations of success.

The pressing challenges of our time — clean energy generation and storage, sustainable chemical processes and precision therapeutics — demand scientific communities integrating technical innovation with chemical understanding. The current divide hinders our ability to address urgent problems. By rebuilding scientific communities with integration as their organizing principle, we can begin to leverage the full potential of AI for chemical discovery.

N. M. Anoop Krishnan [1,2,7] ✉ **& Kevin Maik Jablonka** [3,4,5,6,7] ✉

[1]Department of Civil Engineering, Indian Institute of Technology Delhi, New Delhi, India. [2]Yardi School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi, India. [3]Laboratory of Organic and Macromolecular Chemistry, Friedrich Schiller University Jena, Jena, Germany. [4]Center for Energy and Environmental Chemistry Jena, Friedrich Schiller University Jena, Jena, Germany. [5]Jena Center for Soft Matter, Friedrich Schiller University Jena, Jena, Germany. [6]Helmholtz Institute for Polymers in Energy Applications Jena, Friedrich Schiller University Jena, Jena, Germany. [7]These authors contributed equally: N. M. Anoop Krishnan, Kevin Maik Jablonka. The ordering of the authors was determined by coin toss.

✉e-mail: krishnan@iitd.ac.in; mail@kjablonka.com

Published online: 14 August 2025

References

1. Cheetham, A. K. & Seshadri, R. Artificial intelligence driving materials discovery? Perspective on the article: Scaling deep learning for materials discovery. *Chem. Mater.* **36**, 3490–3495 (2024).
2. Lowe, D. Robots queuing up to fail. *Chemistry World* https://www.chemistryworld.com/opinion/robots-queuing-up-to-fail/4020705.article (2025).
3. McGreivy, N. & Hakim, A. Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations. *Nat. Mach. Intel.* **6**, 1256–1269 (2024).
4. Narayanan, A. & Kapoor, S. Why an overreliance on AI-driven modelling is bad for science. *Nature* **640**, 312–314 (2025).
5. Riebesell, J. et al. A framework to evaluate machine learning crystal stability predictions. *Nat. Mach. Intel.* **7**, 836–847 (2023).
6. Fu, X. et al. Forces are not enough: benchmark and critical evaluation for machine learning force fields with molecular simulations. *Trans. Mach. Learn. Res.* https://openreview.net/forum?id=A8pqQipwkt (2023).
7. Harris, C. et al. Benchmarking generated poses: how rational is structure-based drug design with generative models? Preprint at https://doi.org/10.48550/arXiv.2308.07413 (2023).
8. Mahmood, F. A benchmarking crisis in biomedical machine learning. *Nat. Med.* **31**, 1060 (2025).
9. Bates, S., Jordan, M. I., Sklar, M. & Soloff, J. A. Principal-agent hypothesis testing. Preprint at https://doi.org/10.48550/arXiv.2205.06812 (2024).
10. Sarewitz, D. Saving science. *The New Atlantis* https://www.thenewatlantis.com/publications/saving-science (2016).

Acknowledgements

K.M.J. thanks the Carl-Zeiss Stiftung for the support of this work. N.M.A.K. acknowledges the funding support by the Alexander von Humboldt Stiftung. The authors also thank their research groups (in particular, N. Alampara and M. Ríos-García) and numerous colleagues for discussions.

Competing interests

The authors declare no competing interests.