https://doi.org/10.1038/s43588-023-00446-x

# The promise and pitfalls of AI for molecular and materials synthesis

Nicholas David, Wenhao Sun & Connor W. Coley



As artificial intelligence (AI) proliferates, synthetic chemistry stands to benefit from its progress. Despite hidden variables and 'unknown unknowns' in datasets that may impede the realization of a digital twin for the laboratory flask, there are many opportunities to leverage AI and large datasets to advance synthesis science.

There is an ambition that unsolved problems in science and engineering might be accessible to artificial intelligence (AI) if provided enough empirical data to train on, which could then perhaps make accurate judgements on problems that humans struggle to resolve. The field of synthetic chemistry offers many such problems. Given the past century of modern chemistry, our collective understanding of synthetic chemistry and the empirical outcomes of chemical reactions are distributed across journals, patents, and notebooks. Subsets of this institutional knowledge have been processed into large-scale reaction databases, so there is a tantalizing opportunity to use these datasets in combination with AI to drive synthesis science and supplant heuristics and trial-and-error synthetic chemistry.

Just as decision-making algorithms represent a major AI challenge in self-driving cars, synthesis planning algorithms are a primary bottleneck in realizing self-driving chemical synthesis laboratories[1–3]. Even if we know what hypothetical molecule or material we would like to synthesize and have the hardware to do so, designing a detailed process with full specification of reaction conditions is a formidable challenge. Here, we reflect upon the available datasets of both organic and inorganic chemical syntheses and the capabilities of models available today. We chart a path towards how our community might realize the promise and potential of AI for synthesis science (Fig. 1). We believe that hidden variables of synthesis representing the 'unknown unknowns' in our training dataset are a particularly urgent issue to address. Nevertheless, there are exciting and valuable applications of machine learning (ML)/AI as a 'laboratory assistant' that can process experimental observations and suggest new directions to explore.

## Short-term vision: accessible and searchable data

Collecting and categorizing existing knowledge into structured databases offers substantial added-value over traditional search engine approaches to literature review. The digitization of data enables programmatic queries to answer questions like, "What has been tried before?" and perhaps more importantly, "What has not been tried before?"

For synthetic organic chemistry, such databases have existed for many years in the form of commercial offerings such as SciFinder (www.scifinder.cas.org) and Reaxys (www.reaxys.com), and more recent reaction collections such as Pistachio (www.nextmovesoftware.com), the open access USPTO dataset[4], and the emerging Open Reaction Database[5]. Additional information – exempt from the literature bias towards high-yielding reactions – exists in firewalled electronic notebooks. These resources make experimental records accessible in a digital format and are routinely used to identify known synthetic preparations to molecules of interest, yet they can only contain as much procedural detail as a researcher originally reported, which may not be sufficient for the field's longer-term goals.

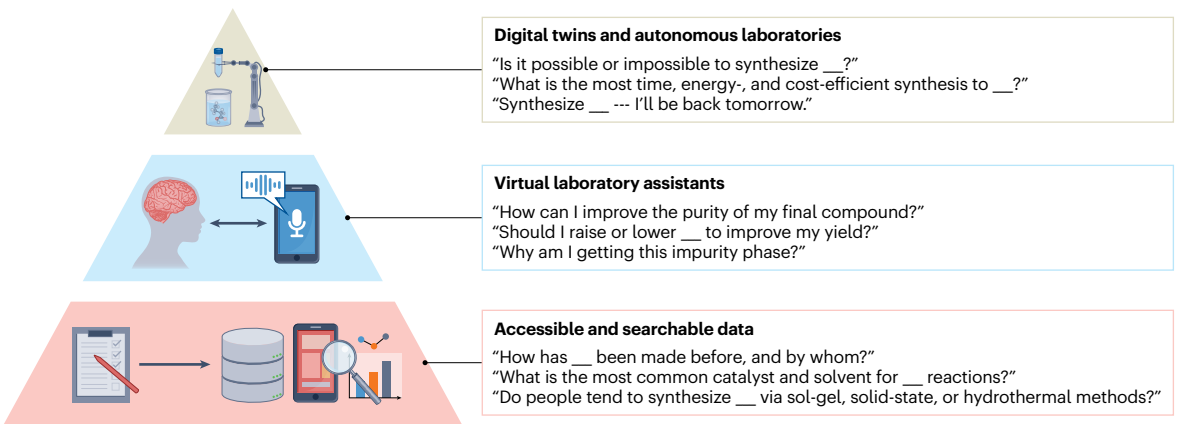


**Fig. 1 | Opportunities in AI for synthesis science in the style of Bloom's taxonomy.** Starting from a foundation of accessible and searchable data (short-term), more advanced capabilities can be built in the form of virtual laboratory assistants (medium-term) as a preface to more robust digital twins and autonomous laboratories (long-term). Example questions scientists might be empowered to ask at each stage are noted.

In inorganic materials chemistry, no commercial databases that catalog reaction recipes and synthesis outcomes exist. However, natural language processing algorithms can be applied to the corpus of scientific data to learn semantic relationships[6,7] and to prepare large, structured datasets. Successful examples include the text-mined dataset of inorganic materials synthesis reactions, which provides 19,488 inorganic solid-state synthesis recipes[8] and 35,675 solution-based (hydrothermal and solvothermal) synthesis recipes[9]. Even though the ~50,000 machine-classified reactions from these examples are far more than what have been cataloged before, approximately only 1 out of 4 text-mined paragraphs could be fully processed by the ML pipeline. This could be improved by standardizing the language of synthesis methods and validating procedures before publication (www.syncheck.org)[10].

Even without machine learning, data visualizations such as histograms and distribution maps can be instructive to a scientist — identifying empty places in parameter space or chemical space to screen. Data visualization can elucidate trends in the general conditions for the synthesis of Li-ion batteries[8], or provide 'cheat-sheets' for specific reaction types (for instance, Buchwald–Hartwig coupling[11]).

## Medium-term vision: virtual laboratory assistants

While searchable synthesis databases enable a more powerful 'recall' approach to synthesis planning than using online search engines, AI can and should be able to provide higher-level analysis, reasoning, and abstraction. If building databases is like having access to every scientist's laboratory notebook, the next goal is having access to every scientist's brain.

There is an opportunity for AI-based laboratory assistants to facilitate creativity by suggesting a new pathway that had not previously been considered by a human chemist operating within their comfort zone. At a high level, this might look like a recommendation system that proposes what procedures or precursors to try and — in response to newly observed experimental data — what to try next. Making creative recommendations would require some sort of novelty metric and a mechanism for quantifying dissimilarity in synthetic pathways, perhaps analogous to metrics from natural language processing word embeddings[12].

Retrosynthesis tools for organic synthesis are one example where one can already see value in idea generation and a collaborative partnership between AI and expert chemists[13]. In the field of inorganic synthesis, the concept of retrosynthesis differs significantly, as syntheses often are one-pot, exhibit high sensitivity to differences in mechanical processing, and do not involve isolating intermediate structures as 'checkpoints' that provide opportunities for control. A natural approach to a challenging materials preparation would be to draw analogies to previously successful reactions (at the expense of creativity). For an AI system to recapitulate that strategy, it must be able to quantify the synthetic similarity between a target material and known materials from its knowledge base of existing recipes. The nature of this similarity metric may involve elemental similarity, or structural similarity, though it is not clear that this translates to synthesis similarity.

To achieve the highest value, AI recommendations must be consistent enough with existing knowledge to have a reasonable chance of success. Responding to unexpected synthesis outcomes should take a holistic view of reactivity and not only follow established sequential experimental design techniques (for example, black box Bayesian optimization[14]) to adjust parameters around a local optimization target. Achieving this vision may require the development of models with stronger priors informed by chemistry, thermodynamics and kinetics, rather than domain-agnostic models.

## Long-term vision: digital twins and autonomous laboratories

The long-term vision for predictive chemical synthesis is akin to a full digital twin — a virtual representation of a synthetic process that can predict the outcome of a reaction and optimize conditions prior to any physical experimentation. This would substantially reduce the need for costly and time-consuming physical trials by transferring the burden of condition/parameter optimization from experimental to digital exploration. Expanding this vision, connecting such a digital twin with a robotic laboratory would enable a fully closed-loop autonomous 'self-driving' laboratory, which, given a desired target material, could fully execute a synthesis from start to finish[1–3,15].

The challenge is that such a digital twin, were it to exist, needs essentially perfect predictive accuracy. We often assume that the experimental data on which models are trained represents some unarguable 'ground truth' about chemical reactivity and about the mapping of synthetic recipes to product identity, yield, rate, and so forth. But it is well known that the outcome that one chemist observes often does not match what is observed by another. Uncertainties or ambiguities in synthesis procedures obfuscate subtler patterns driving reactivity, limiting the accuracy and generalizability of the models that we train.

Since many of these confounding variables represent 'unknown unknowns' in synthesis datasets, they may represent fundamental impediments towards training a perfectly predictive synthesis digital twin. Consider the example of the reported (now retracted) metal-free Suzuki coupling[16], describing high-yielding C–C formation without the use of a palladium catalyst, which despite extensive efforts could not be reproduced in others' labs. The successful coupling observed by the authors was later explained by impurities[17] whose existence was not known at the time of publication. In another example, during the synthesis of single crystal *o*-$Nb_{12}O_{29}$ and $Nb_{12}O_{54}$, 2 mol% rhodium was added to "aid crystallization" of the target materials[18]. No understanding or explanation of the role that rhodium played in crystallization was given. To anticipate the outcomes in these two examples, a digital twin would need to (a) be aware of the possibility of confounding impurities and (b) recognize how the absence of an additive will impact performance even if there is no mechanistic understanding or prior evidence of that additive's influence. Without extensive characterization of starting materials and equipment or more rigorous reporting of ablation studies, this remains a long-term prospect.

Many other aspects of synthetic procedures are also underreported despite recognition of their importance: order of addition, ambient temperature/humidity (not just 'room temperature'), use of glovebox or Schlenk line (not just 'inert atmosphere'), and vendors and lot numbers of purchased reagents. Certain protocols involving bespoke electrochemical or photochemical setups are notoriously difficult to reproduce exactly, though standardization is improving with newer commercial systems. Similar uncharacterized aspects of reactions contribute to scale-up challenges depending on heat/mass transfer limitations, as a reported yield may be a function of the process as much as the reagents involved. Most database efforts are directed towards structured representations of reactions using well-defined schemas. However, the peculiarities of a particular synthesis — especially if inorganic — may be difficult to include without unstructured fields for additional process notes.

To address this, we need to change the way the chemistry community presents and stores reaction information. Oftentimes, it is only the final successful synthesis methodology that is reported; moreover, it is reported in sparse and minimal prose. There is broad agreement that reporting failed attempts[19] helps us (and machines) learn more synthesis science and that understanding the 'secret sauce' is essential for long-term reproducibility of experimental methods. High-throughput laboratories present an opportunity to directly catalog all reaction inputs and outputs from experiments — including failed 'dark reactions'. Reporting negative outcomes will require a culture shift, but it is an essential one if we want synthesis to be more predictable in the future.

## Beyond structured data: uncovering hidden opportunities

There is a rich dataset of implicit information beneath published synthesis papers that could be mined through clever ways beyond supervised learning. For example, a synthesis recipe with long reaction times (>2 weeks), unusual precursors, or laborious synthesis routes implies that more straightforward approaches were probably unsuccessful. Such cases can be examined more thoroughly afterwards from a theoretical perspective. This sort of implied information can also be used to augment existing reaction datasets, for example, by viewing a published reaction of A + B → C as evidence that A + B does not react to form D, E, F, and so forth under like conditions[20].

An additional application of large-scale datasets is that they can enable retroactive experiments to test new hypotheses. We recently hypothesized that in aqueous synthesis, the kinetics of structure selectivity can be informed by the maximal thermodynamic driving force as calculated from the Pourbaix free energy[21]. Using the text-mined solution synthesis dataset[9], this hypothesis could retroactively explain the empirically observed synthesis conditions for a large number of reactions. Even in cases where hypothesis generation is handled by creative human scientists, validation can make use of the empirical dataset.

## Outlook

Achieving the ultimate goal of a synthesis AI digital twin is a long-term goal of synthetic chemistry and is not the burden of experimental or computational experts alone. To begin to eliminate confounding unknowns, there should be a broader adoption of in-house reproducibility tests, where another member of the lab, or another lab, repeats an experiment based solely on a written procedure prior to publication. The origin and properties of starting materials are anecdotally the most heavily cited reason for reproducibility failures: vendor versus in-house synthesis, purity/quality, and repurification after purchase. Prose describing these factors can be replaced with a defined data structure, such as the authors' own Open Reaction Database format for organic reactions[5]. But subsequent steps will need to evolve the schema to accommodate additional information and to evolve our learning algorithms as we learn how to best make use of it. We echo the reproducibility concerns of Bergman and Danheiser[22] not only for the sake of good science, but also for building new AI-driven tools for tackling organic and inorganic materials synthesis.

While our community works towards this long-term goal, computational scientists have many exciting opportunities to build AI-driven recommendation engines that can aid in making creative and inspiring suggestions to guide new synthesis strategies. This will require algorithms to achieve higher-level understanding of chemistry–structure–synthesis relationships. Such a problem is scientific in nature and not just software, but data (both existing results and new experiments) will play an essential role in resolving these relationships. More generally, we recommend looking beyond the traditional (and now obvious) path of doing direct supervised machine learning to propose synthesis recipes. Considering how to propose new hypotheses manually or algorithmically and evaluate them against a dataset is a complementary and promising, interesting path to improve synthesis science.

**Nicholas David [1], Wenhao Sun [1] ✉ & Connor W. Coley [2,3] ✉**
[1]Department of Materials Science & Engineering, University of Michigan, Ann Arbor, MI, USA. [2]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA.
✉e-mail: whsun@umich.edu; ccoley@mit.edu

Published online: 18 May 2023

## References

1. Burger, B. et al. *Nature* **583**, 237–241 (2020).
2. Coley, C. W. et al. *Science* **365**, eaax1566 (2019).
3. Abolhasani, M. & Kumacheva, E. *Nat. Synth.* https://doi.org/10.1038/s44160-022-00231-0 (2023).
4. Lowe, D. *Chemical Reactions From US Patents (1976-Sep2016)* (Figshare, 2017); https://doi.org/10.6084/m9.figshare.5104873.v1
5. Kearnes, S. M. et al. *J. Am. Chem. Soc.* **143**, 18820–18826 (2021).
6. Tshitoyan, V. et al. *Nature* **571**, 95–98 (2019).
7. Olivetti, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
8. Kononova, O. et al. *Sci. Data* **6**, 203 (2019).
9. Wang, Z. et al. *Sci. Data* **9**, 231 (2022).
10. Kim, E., Huang, K., Kononova, O., Ceder, G. & Olivetti, E. *Matter* **1**, 8–12 (2019).
11. Fitzner, M. et al. *Chem. Sci.* **11**, 13085–13093 (2020).
12. Kim, E. et al. *Chem. Mater.* **29**, 9436–9444 (2017).
13. Segler, M. H. S., Preuss, M. & Waller, M. P. *Nature* **555**, 604–610 (2018).
14. Szymanski, N. J. et al. *Mater. Horiz.* **8**, 2169–2198 (2021).
15. MacLeod, B. P. et al. *Sci. Adv.* **6**, eaaz8867 (2020).
16. Xu, L. et al. *Nat. Catal.* **4**, 71–78 (2021).
17. Novák, Z. et al. *Nat. Catal.* **4**, 991–993 (2021).
18. McQueen, T., Xu, Q., Andersen, E. N., Zandbergen, H. W. & Cava, R. J. *J. Solid State Chem.* **180**, 2864–2870 (2007).
19. Raccuglia, P. et al. *Nature* **533**, 73–76 (2016).
20. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. *ACS Cent. Sci.* **3**, 434–443 (2017).
21. Wang, Z. et al. Optimal thermodynamic conditions to minimize kinetic byproducts in aqueous materials synthesis. Preprint at https://doi.org/10.21203/rs.3.rs-2398824/v1 (2023).
22. Bergman, R. G. & Danheiser, R. L. *Angew. Chem. Int. Ed.* **55**, 12548–12549 (2016).

## Acknowledgements

N.D. and W.S. acknowledge support through the Dreyfus Program for Machine Learning in the Chemical Sciences and Engineering. C.W.C. thanks the National Science Foundation under grant no. CHE-2144153 and the AI2050 program at Schmidt Futures (grant G-22-64475) for financial support. We thank P.F. Poudeu, J. Neilson, and A. Miura for stimulating conversations.

## Author contributions

N.D., W.S. and C.W.C. jointly prepared the manuscript.

## Competing interests

The authors declare no competing interests.