DATA-DRIVEN MATERIALS SCIENCE: FOR A LAB OF THE FUTURE

JACQUELINE M. COLE[*]

[*] *Ray Dolby Centre, Cavendish Laboratory, Department of Physics, University of Cambridge,*
J. J. Thomson Avenue, Cambridge, CB3 0HE, United Kingdom

AI-driven materials discovery has captured the world's imagination with software giants and world-leading research labs claiming or contending all sorts of breakthroughs. that accelerate materials science. AI-driven efforts that use computational methods to predict materials with targeted physical or chemical properties are of particular contention. Only last week was this debate amplified by a contentiously titled *Nature* article: "AI is dreaming up millions of new materials. Are they any good?"[1].

There are clearly many challenges to overcome and yet opportunities to harness. This viewpoint article attempts to address the current hype in AI for materials discovery, by showcasing the current status of the field; and focusing on using experimental data, rather than computed data, in concert with AI, to present results from real-world data.

Challenges and Opportunities in AI for Materials Discovery

What actually makes AI intelligent in materials science? I'll argue for three key things based on my personal perspective:

(1) Data

AI methods are useless without data to train them. More specifically, AI algorithms need to be trained on sufficient, relevant data.

Data quantity: Each type of AI algorithm can be described by a certain number of parameters that need to be optimised with enough relevant data to produce a reliably predictive AI model. A key consideration is the data : parameter ratio which should be no less than 10 : 1 to be minimally viable from a statistical perspective.

Data types - Chemical space: As well as raw data quantity, one should also consider how balanced is a dataset. In the field of materials science, an AI model will best serve a given chemical problem if it has been trained on chemical data that relate to that problem as well as a broader range of chemical data.

Data types - Chemical data on positive, null and negative results. Datasets that describe how chemicals respond to external effects (e.g. chemical reactions, chemical properties) should ideally contain a mixture of positive, null or negative responses to ensure a balanced dataset. However, with few exceptions, such data nearly always comprise positive results in the bulk or entirety of the dataset. High-throughput experiments (HTEs) are one such exception as they record a range of measurements that typically show positive, null or negative types of data.[2] Likewise, electronic laboratory notebooks that log positive as well as null or negative research results are a source of balanced data.

Data types - Experimental data. These depict real-world chemical data which are much more heterogeneous than computed data by virtue of chemicals being affected by all sorts of experimental environments, e.g. device environment, interfacial effects; and containing various types of experimental error.[3] In stark contrast, computed chemical data are much smoother, being exact e.g. from calculations on isolated molecules, and lacking in any experimental error (albeit they may manifest systematic computational error); moreover, calculations can easily map out an entire region of chemical space, subject to the availability of sufficient computational resources. Such complete mappings have not tended to be possible experimentally until recently with the latest advances in robotic autonomous laboratories, enabling HTEs that offer such mapping options.

(2) Representations

Aside from raw data, the intelligence of AI for materials science lies in encoded forms of relationships between chemicals and their responses e.g. chemical-reaction patterns or structure-property relationships; these issue patterns in the data that machine-learning (ML) algorithms can identify and learn to correct predict chemical reactions or predict materials with designer functionality.

Representations comprise the way in which these relationships have been encoded via custom algorithms that depict systematic patterns in chemical reactions or known structure-property relationships. The smarter representations tend to be more comprehensive and efficient in their description of the chemical relationships involved.

(3) Design-to-device pipelines for materials discovery.

Supply-chain management for materials science is needed if AI is to be used for materials discovery. A bespoke pipeline is required to solve a given type of materials problem. For example, a data-driven materials discovery pipeline[4] may manifest as:

Data Source > Data Mine > Predict properties > experimentally validate

An autonomous lab using HTEs that test a series of ligands, l, may use:

Reactant > Add a ligand, l > Characterise product > Measure properties

My provision of large experimental datasets for materials research

We released the world's first 'chemistry-aware' natural-language-processing tool, ChemDataExtractor,[5,6] for the materials-research community. This software enables scientists to automatically mine text from scientific documents (e.g. the academic literature) to auto-generate large custom materials datasets to meet the bespoke data needs of a given application of interest. Thereby, experimental datasets can be auto-curated to train ML models for materials science.

Popular examples of experimental datasets that have been created by ChemDataExtractor users include "DigiMOF" on metal-organic frameworks[7] and a dataset on self-cleaning coating materials[8].

The Cole group has also used ChemDataExtractor to auto-generate 12 open-source experimental datasets to serve the materials community; each contain 20k-720k data records, culminating in a total of 2.15 billion data records for materials science. Each dataset covers a distinct materials-science domain. The stress-strain engineering,[9] battery[10] and photovoltaics[11] datasets are the world's largest auto-generated experimental materials datasets, comprising ca. ¾, ¼ and ½ million records of chemicals and materials or device properties, respectively. Such sizes meet the data needs of some of the most sophisticated machine-learning (ML) algorithms.

Thereby, we have trained materials-domain-specific language models for stress-strain engineering[12], batteries[13], photovoltaics[14], and other domains, using corpora sourced from the experimental datasets.

ChemDataExtractor has also been employed to create large question-answering datasets that can be used to fine-tune these materials-domain-specific language models for prompt engineering.[14,15]

Our ChemDataExtractor-generated databases have been used in a 'design-to-device' pipeline to afford data-driven materials discovery;[4] the most successful exemplar being the discovery of new light-harvesters for dye-sensitized solar cells.[16]

Digital assistants for autonomous labs in materials science

Supply-chain management of chemistry-informed AI pipelines is crucial for operating autonomous robotic laboratories. The realisation of such AI pipelines has reached the demonstration phase via the AI agent, *ChemCrow*, which is based on a large-language model.[17] Major improvements in such agentic-AI demonstrations are needed before any digital-lab assistant for chemistry can operate with a sufficiently high quality that it is practically viable within a human-AI user mode; let alone one that is usable as an autonomous lab agent.

Notwithstanding such current limitations in the state-of-the-art, some parts of the supply chain for data-driven materials discovery can be automated with AI. For example, AI-based materials characterisation is now possible. Thereby, Cole *et al*. were the first to demonstrate how an AI model could automatically identify the molecular structure of a material directly and solely from a raw infra-red spectral image.[18] Their AI model was based on a convolutional neural network (CNN) that had been trained on 50,000 experimental data. Its CNN architecture lay the foundations for the automatic materials characterisation of any type of raw diffraction or spectroscopy data. Indeed, Liu and Cole have just reported a CNN that can automatically classify ^{1}H or ^{13}C nuclear magnetic resonance (NMR) spectra from chemical solutions.[19]

Cole *et al.* also built a transformer-based neural-network architecture to predict nanostructural shape and size from small-angle X-ray scattering data, albeit using computational data to train their AI-model.[20] They highlight the need for experimental data to better train their AI models. This AI need for such data runs throughout this paper.

References

1. M. Peplow, *Nature* **646**, 22 (2025).
2. B. Burger, P. M. Maffettone, V. V. Gusev, C. M. Aitchison, Y. Bai *et al.*, *Nature* **583**, 237 (2020).
3. F. Strieth-Kalthoff, F. Sandfort, M. Kuehnemund, F. R. Schaefer, H. Kuchen, *et al.*, *Angew. Chemie Int. Ed.* **61**, e202204647 (2022).
4. J. M. Cole, *Acc. Chem. Res.* **53**, 599 (2020).
5. M. C. Swain, J. M. Cole, *J. Chem. Inf. Model.* **56**, 1894 (2016).
6. J. Mavracic, C. J. Court, T. Isazawa, S. R. Elliott, J. M. Cole, *J. Chem. Inf. Model.* **61**, 4280 (2021).
7. L. T. Glasby, K. Gubsch, R. Bence, R. Oktavian, K. Isoko *et al.*, *Chem. Mater.* **35**, 4510 (2023).
8. S. Wang, Y. Wan, N. Song, Y. Liu, T. Xie *et al.*, *Sci. Data* **11**, 146 (2024).
9. P. Kumar, S. Kabra, J. M. Cole, *Sci. Data* **11**, 1273 (2024).
10. S. Huang, J. M. Cole, *Sci. Data* **11**, 1273 (2024).
11. E. J. Beard, J. M. Cole, *Sci. Data* **9**, 329 (2022).
12. P. Kumar, S. Kabra, J. M. Cole, *J. Chem. Inf. Model.* **65**, 1873 (2025).
13. S. Huang, J. M. Cole, *J. Chem. Inf. Model.* **62**, 6365 (2022).
14. Z. Li, J. M. Cole, *Digital Discovery* **4**, 998 (2025).
15. O. Sierepeklis, J. M. Cole, *J. Chem. Inf. Model.* **65**, 8579 (2025).
16. C. B. Cooper, E. J. Beard, A. Vazquez-Mayagoitia, L. Stan, G. B. G. Stenning *et al.*, *Adv. Energy Mater.* **9**, 1802820 (2019).
17. A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White *et al.*, *Nat. Mach. Intell.* **6**, 525 (2024).
18. G. J. Jung, S. G. Jung, J. M. Cole, *Chem. Sci.* **14**, 3600 (2023).
19. S. Liu, J. M. Cole, *J. Chem. Inf. Model.* **65**, 8435 (2025).
20. B. Yildirim, J. Doutch, J. M. Cole, *Digital Discovery* **3**, 694 (2024).