https://doi.org/10.1038/s41592-025-02995-7

The missing data for intelligent scientific instruments

Henry Pinkard & Nils Norlin

Most scientific instruments currently discard rich streams of commands, data and metadata from which AI systems could learn to conduct experiments with expert-level decision-making and troubleshooting skills. Recording and using this data at scale requires rethinking what data to store, incentivizing large-scale cooperation, and determining how to quantify the reliability of such autonomous systems.

The failure to preserve researchers' expertise in digital form results in a missed opportunity to create highly capable scientific instruments that accelerate scientific progress and enhance reproducibility. Every day, microscopes, flow cytometers, spectrometers and other instruments across thousands of laboratories produce rich digital traces of experimental work: precise sequences of commands, real-time observations and moment-by-moment decisions that embody years of hard-won expertise by their operators. At the same time, artificial intelligence (AI) has reached unprecedented capabilities in learning from structured data. Modern transformer neural networks[1] can learn statistical relationships across domains from language[2] to proteins[3] and robotic actions[4]. This has already enabled autonomous experimental systems[5], for example, in chemistry[6] and materials science[7], where extensive data are available. We believe that enabling AI systems that can carry out complete biological experiments across a variety of instruments could disseminate experimental expertise broadly across the scientific community.

Currently, the primary way we transmit researchers' expertise remains through written methods sections, which compress complex sequences of expert work into brief summaries. These summaries do not contain key knowledge and skills needed for reproducibility, such as the ability to recognize problems, optimize parameters and adapt procedures in real-time. To learn these skills, researchers often must visit specific laboratories or attend extensive training courses, and when an experienced researcher leaves a lab, their expertise often goes with them.

At this moment, most automated systems cannot perform general-purpose, adaptive decision making. For example, current smart microscopy systems[8] rely on hard-coded rules or are designed for specific situations. Although they can already operate at scales and speeds beyond human capabilities, they lack the ability to plan, troubleshoot and adapt to unexpected real-time observations[9].

Despite the limitations of current automated systems, AI technologies show potential, particularly when trained on larger, higher-quality datasets. Recent AI breakthroughs in scientific applications have consistently emerged in domains with extensive data availability[10], reflecting the predictable relationship between transformer performance and data volume[2]. This pattern suggests that AI systems could potentially learn the adaptive decision-making capabilities that current automation lacks. To take advantage of these scaling relationships for scientific applications, AI development could benefit from new sources of domain-specific, high-quality training data. Such data would be particularly valuable for training since the conventional source, publicly available text, is approaching exhaustion[11], thereby limiting further progress along traditional scaling curves.

We propose that systematically capturing digital traces of expert work could transform scientific instruments from rule-based automation systems into intelligent research partners. By recording complete sequences of commands, observations and real-time decisions across multiple experiments, instruments and operators, we could train AI systems that surpass current hard-coded automation. Such systems could progress from copilot-style assistance to fully autonomous agents capable of planning, troubleshooting and real-time adaptation.

Beyond enabling more capable instruments, we envision this approach transforming how scientific expertise spreads, allowing researchers to rapidly adopt new methodologies while providing a rigorous new standard for reproducibility.

To illustrate how workflow data could enhance scientific instruments, we examine a concrete example: a transformer neural network that learns to focus a microscope (Fig. 1). This example illustrates the broader potential for AI to learn from the rich streams of data generated during scientific experiments.

Transformers: a common tool in modern AI

Transformer-based systems excel at finding patterns in sequential, heterogeneous data, such as that produced by instruments during experiments. Unlike traditional automation that requires explicit programming for each scenario, transformers learn by discovering relationships between elements in their input, enabling them to process diverse data types, from commands to images, all within a unified framework.

Many transformers (for example, GPT-3 and GPT-4) use a 'decoder-only' autoregressive architecture, which predicts the next 'token' in a sequence given the previous ones. Tokens are fundamental units of data: characters or words in language, pixels or patches in images[12]. Figure 1a illustrates this next-token prediction process, where a transformer model repeatedly predicts each subsequent token on the basis of the accumulated context, gradually building up complex sequences.

Training these models on large datasets creates rich internal representations, whose knowledge can be elicited through fine-tuning to perform specific tasks. The resulting systems can perform various applications, such as engaging in conversations or suggesting code completions.

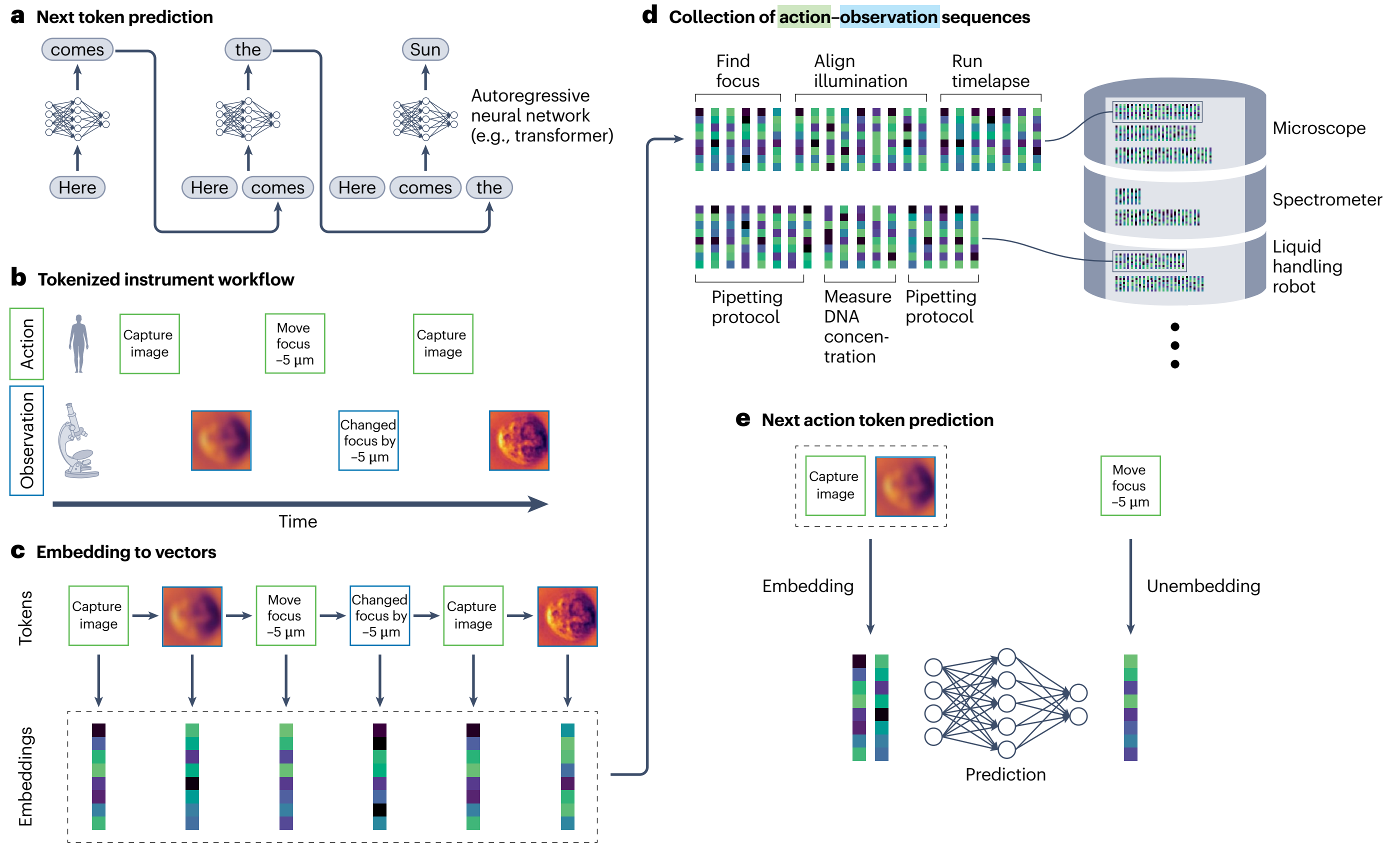

Fig. 1 | Training an autoregressive model from scientific instrument usage. **a**, Autoregressive models predict the next token in a sequence on the basis of previous context. **b**, Scientific instruments naturally generate structured sequences through the interaction between operator and instrument – operators take actions (green) and the instrument responds with 'observations' consisting of data and state changes (blue). **c**, To enable learning from these heterogeneous data types (commands, images, metadata), all tokens are embedded into vectors. **d**, A database of these vector sequences from many experiments captures the diverse ways that instruments are used to accomplish experimental tasks. **e**, Example of action prediction: given an out-of-focus image, the model can predict the appropriate stage movements needed to achieve focus, just as it could learn to predict actions for more complex experimental tasks.

Neural networks derive their power from transforming raw data into learned representations – numerical encodings that capture key relationships. As information flows through the model's layers, these representations become increasingly context-aware and abstract, providing a way to process different types of data while preserving meaningful relationships between concepts.

The model's capabilities scale with its complexity: more sophisticated models can recognize nuanced features such as specific behaviors or contexts, encoding these distinctions within different parts of their vector space.

Access to these representations simplifies complex recognition tasks. Training a vision system to identify, for example, fibroblasts, would require relatively few examples when built on existing embeddings. This is in stark contrast to writing explicit rules for every variation in shape, illumination and cell subtype, or having to learn such rules from scratch.

From language to action. Recent advances in vision–language–action models[4] demonstrate how AI systems can successfully move beyond language and vision to control physical systems. These systems leverage pretrained vision and language models while encoding physical actions as text tokens that can be converted into precise control commands. Unlike in robotics applications that must bridge the physical–digital gap through complex demonstrations or teleoperation, many scientific instruments are already operated digitally. Their software interfaces already generate structured data during normal operation (for example, commands, responses and state changes), making them ideal candidates for AI control.

Tokenizing scientific workflows. Scientific instrument control can be framed as a sequence modeling problem, just like language generation. The digital nature of these instruments means that controlling them is much like speaking with a chatbot, only with different data types: rather than text being tokenized for questions and answers, commands and instrument responses (data and metadata) can be converted into their respective tokens. Figure 1b illustrates this structure in microscopy, where an operator issues commands such as "capture image" and "move focus" (normally triggering a series of low-level API calls) and the instrument responds with images and text descriptions of state changes.

These sequences of actions, observations and metadata tokens collectively capture information about the instrument's state, the experimental progress and facets of the operator's intent. This information is embedded in the statistical relationships between tokens, much as letters combine into words and words into meaningful phrases to capture complex meaning in text. Through these relationships, tokens can be composed to represent experimental tasks of increasing complexity, from basic instrument operations to sophisticated experimental protocols.

Existing transformer architectures provide a direct path to capturing and using this knowledge of experimental execution. All tokens are first embedded as vectors (Fig. 1c) and then collected in a database, ideally including a diversity of experimental action sequences across different operators, instruments and samples (Fig. 1d). Training on these sequences enables models to learn patterns of experimental execution, just as language models learn to predict the next word in a sentence.

Figure 1e illustrates a simple example: given an out-of-focus image, the model learns to generate appropriate stage movements to achieve focus, just as a human operator would. This capability would naturally generalize across instruments; learning to focus one microscope is quite similar to focusing others.

In our view, the most efficient path to implementing these systems is likely to be to leverage existing frontier models rather than train new ones from scratch. Pretrained vision models could process experimental images while language models handle commands and metadata, creating powerful systems with relatively small instrument-specific training datasets. This approach would capture the essential patterns of experimental workflows while minimizing both data requirements and computational costs of training.

Open challenges

While the potential for learning from experimental workflows is clear, we expect that several practical challenges must be addressed before this vision can be realized. These range from technical hurdles around data volumes and infrastructure to social challenges around incentives and reliability standards.

Storing and sharing data. Experiments can routinely generate substantially more data than goes into final analyzed results, making complete workflow capture impractical. Smart strategies for downsampling and selecting the most informative interactions will be crucial for making this approach tractable. Effective data sharing presents an additional challenge. The greatest benefits will emerge when many laboratories pool data together for training, but this requires technical standards, data repositories, community guidelines[13] and new incentive structures, for example, for data curation.

Software limitations. Most existing scientific instruments were not designed with comprehensive data capture in mind and typically save only the final results, not the complete workflows that produced them. However, emerging software frameworks that wrap existing device control systems could enable workflow capture without rebuilding software infrastructure from scratch (https://exengine.readthedocs.io/en/latest/).

Incomplete digitalization. Furthermore, many life science workflows still include manual steps outside digital systems. However, an increasing number of devices syncing with electronic lab notebooks, digitalized pipettes and automated liquid handlers are creating increasingly complete digital records. This digitization is likely to accelerate[14], incentivized by the potential of autonomous workflow control.

Decentralized AI training. While many contemporary frontier models are trained with thousands of high-end GPUs running for months at a time, algorithmic improvements continue to lessen these requirements. Rather than requiring massive centralized GPU clusters, emerging approaches[15] suggest that highly capable AI models could be trained across networks of modest laboratory computers. This would shift the challenge from acquiring expensive computing resources to coordinating access to existing infrastructures and even personal computers. Future systems might allow scientists to access models in proportion to their contributed resources (data and computing) or offer ways to establish citizen science projects for AI development. Such implementations are likely to require new incentive structures, technical standards, and support from funding agencies and scientific institutions.

Reliability. Modern AI systems, while powerful, can produce hallucination-based errors that pose unique challenges for scientific applications. Hallucinations cannot be completely prevented in current deep learning systems, making empirical quality control essential for scientific instruments that incorporate AI. Practical approaches to address this include developing benchmarks that test performance across different experimental scenarios, implementing consistency checks to validate results against known physical constraints, and maintaining oversight mechanisms for system decisions during operation. A graduated deployment – that is, starting with simple tasks and expanding autonomy only as reliability is demonstrated – will allow researchers to build trust while learning the system's capabilities and limitations.

Outlook

Regardless of which AI approach ultimately proves the most effective, the quality and comprehensiveness of the training data are likely to remain the fundamental determinant of system performance for the foreseeable future. While future architectural advances will certainly surpass current approaches and mitigate some of the current limitations, collecting rich workflow data now creates lasting value for scientific instrument automation as data quality is likely to continue to constrain the upper bounds of model capabilities. Despite challenges, the promise of AI appears capable of unlocking a new era of accelerated scientific discovery, where the boundaries between human expertise and machine capabilities become increasingly blurred.

Henry Pinkard[1,2] ✉ **& Nils Norlin**[2,3,4] ✉
[1]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA. [2]Department of Experimental Medical Science, Lund University, Lund, Sweden. [3]Nanolund, Lund University, Lund, Sweden. [4]Lund University Bioimaging Centre, Lund University, Lund, Sweden.
✉e-mail: hbp@berkeley.edu; nils.norlin@med.lu.se

Published online: 31 December 2025

References

1. Vaswani, A. et al. Attention is all you need. In *Advances in Neural Information Processing Systems* Vol. 30 (eds. Guyon, I. et al.) https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (Curran, 2017).
2. Henighan, T. et al. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2010.14701 (2020).
3. Jumper, J. et al. *Nature* **596**, 583–589 (2021).
4. Kim, M. J. et al. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2406.09246 (2024).
5. OECD. *Artificial Intelligence in Science: Challenges, Opportunities and the Future of Research* https://www.oecd.org/en/publications/artificial-intelligence-in-science_a8d820bd-en.html (OECD, 2023).
6. Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. *Nature* **624**, 570–578 (2023).

7. Szymanski, N. J. et al. *Nature* **624**, 86–91 (2023).
8. Eisenstein, M. *Nat. Methods* **17**, 1075–1079 (2020).
9. Carpenter, A. E., Cimini, B. A. & Eliceiri, K. W. *Nat. Methods* **20**, 962–964 (2023).
10. Griffin, C., Wallace, D., Mateos-Garcia, J., Schieve, H. & Kohli, P. A new golden age of discovery: seizing the AI for Science opportunity. *AI Policy Perspectives* https://www.aipolicyperspectives.com/p/a-new-golden-age-of-discovery (2024).
11. Villalobos, P. et al. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2211.04325 (2024).
12. Manning, C., Raghavan, P. & Schuetze, H. *Introduction to Information Retrieval* (Cambridge Univ. Press, 2009).
13. Zulueta-Coarasa, T. et al. *Nat. Methods* **22**, 2245–2252 (2025).
14. Skowronek, P., Nawalgaria, A. & Mann, M. Preprint at bioRxiv https://doi.org/10.1101/2025.10.05.680425 (2025).
15. Douillard, A. et al. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2311.08105 (2023).

Acknowledgements

We thank Ivo Vellekoop for discussions.

Author contributions

H.P. and N.N. conceived the project and wrote the manuscript together.

Competing interests

The authors declare no competing interests.