

AI-Integrated Scientific Inquiry: A Practice-Centered Vision for Science Education

Arne Bewersdorff, Matias Rojas, and Xiaoming Zhai

AI4STEM Education Center, University of Georgia, Athens, GA, USA

Corresponding author: Arne Bewersdorff, arne.bewersdorff@uga.edu

Abstract

Artificial intelligence (AI) has become part of scientific inquiry. Scientists use AI to observe and measure phenomena, to identify patterns in data, and to build models. As AI moves into scientific inquiry, it gains relevance for science education: students should learn how AI is changing scientific practices, ideally by engaging in AI-integrated scientific inquiry themselves. How to design such instruction, grounded in authentic scientific practice rather than taught as a standalone topic, remains an open question. In our vision, which we describe in this article, AI is treated as a set of scientific instruments that students use within the scientific practices described by the Next Generation Science Standards. Each instrument is a genuine scientific tool, pedagogically bounded: its controls are simplified while its core scientific function is preserved. The approach has two aims: engaging students in authentic scientific inquiry, and building an understanding of how AI is used in science and where it can mislead (discipline-based AI literacy, DAIL). In the article, we focus on the investigative core of inquiry, namely observing, analyzing, and modeling, and describe one exemplary AI instrument for each: computer vision for observing, clustering for analyzing, and generative modeling for modeling. We argue that every AI instrument in science education should carry a distinct reflection point that prompts critical evaluation of the AI instrument itself. Finally, we describe how agentic AI, operating across the whole inquiry rather than a single practice, could be represented, arguing that students should first build a foundational understanding of scientific inquiry and AI instruments before relying on agentic AI.

Keywords: scientific inquiry, AI literacy, science and engineering practices, artificial intelligence, science education.

1 Introduction

Artificial intelligence (AI) has moved toward the center of scientific work. AI now supports researchers across the natural sciences as they generate hypotheses, design and conduct experiments, and extract meaning from data at scales beyond the reach of any single research team (Jordan & Mitchell, 2015; Wang et al., 2023). This development runs across methods and disciplines: deep neural networks identify and count wild animals in millions of camera-trap photographs (Norouzzadeh et al., 2018). Deep learning predicts three-dimensional protein structures from sequence (Jumper et al., 2021), and generative models design entirely new proteins that do not occur in nature (Watson et al., 2023). Unsupervised methods cluster astronomical and genomic data and reveal patterns that were not anticipated (Wang et al., 2023). In each of these cases, AI does not replace the scientist; it functions as a new kind of instrument, one that must still be aimed, calibrated, and reflected upon.

As AI becomes part of scientific practices, there is a strong argument that it should become part of scientific literacy and of how science is learned (Zhai, 2025). For science teachers, the central question is not whether AI belongs in the classroom, but how to introduce it so that it strengthens and complements the authentic scientific practices students are already learning, rather than bypassing them by fostering AI literacy in isolation.

General AI literacy frameworks describe what every student should understand about AI. Long and Magerko (2020), for example, describe the competencies a citizen needs to recognize and evaluate AI. These frameworks are valuable, but they are not written for the science teacher who wants to teach students how AI actually changes scientific practices. Recent work has begun to identify what

is missing by distinguishing generic AI literacy from discipline-based (also called domain-specific) AI literacy (DAIL), the competencies required to use and evaluate AI within a particular field or discipline (Knoth et al., 2024; Zhai, 2025). Related work has begun to map how AI is reshaping contemporary scientific practice and what this implies for science education (Nehm & Kubsch, 2026), and a first classroom step has introduced AI-based scientific inquiry using a machine-learning tool (Herdliiska & Zhai, 2024). What remains to be articulated is a concrete account of discipline-based AI instruction for science, one anchored in the actual practices and examples of scientific inquiry. We describe a vision and a platform that enacts it, places AI instruments across scientific inquiry, ties each to an authentic scientific practice, and ensures that students reflect on and critically evaluate the output of AI instruments. The platform is a browser-based learning environment that provides one reduced AI instrument for each scientific practice, with its reflection point built in; three of its instruments are shown in Figures 1–3. In this sense, we define an AI instrument as a tool that performs an act of scientific investigation that would ordinarily require human-like intelligence, such as finding patterns, inferring relationships, or generating structures. Such AI instruments lie along a continuum, from statistical learning such as regression and clustering, via deep learned models such as those used in computer vision, to generative AI. The boundary between statistics and machine learning is fluid (Breiman, 2001). Regression, though rooted in classical statistics, is also the usual entry point to supervised learning (Hastie et al., 2009; Russell & Norvig, 2021). What the instruments share is not a common mechanism but the critical evaluation they demand of the student. AI instruments differ, too, in how far their inner workings can be inspected, a point we return to in Section 2.2.

Beyond these AI instruments, more autonomous AI systems, including agentic AI, are becoming more than tools: they can assist researchers across multiple stages of scientific inquiry. The implications of this overarching role, often framed as AI as collaborator, are discussed in a separate subsection (Section 3.5) and in the discussion that follows.

2 A Vision for AI-Integrated Scientific Inquiry

AI is not a new topic to be added “on top” of science education, but a new generation of scientific instruments, and these instruments belong within the scientific practices. The microscope did not turn biology into the study of lenses; it gave biologists a new way to observe, and with it a new question to ask: whether an observed structure is real or an artifact of the preparation. Computer vision, clustering, and generative models can be understood in the same way. The phenomenon under investigation remains what it has always been, such as why leaves differ in shape, how an object falls, or what makes a molecule sweet. AI serves as the instrument rather than the main object of investigation. At the same time, the role of AI in scientific work is something students are meant to learn about.

The two aims map onto scientific literacy (engaging in authentic scientific practices) and discipline-based AI literacy (understanding the AI instruments themselves). Using AI as instruments within scientific practices advances both together. The two aims are not separate lessons: students come to understand an AI instrument by using a pedagogically bounded version that preserves its core scientific function. These remain design claims about how such instruction should work; whether operating the instruments in fact produces the intended learning is an empirical question we return to (Section 4.4).

2.1 Scientific Practices as the Organizing Framework

To keep scientific inquiry at the center, the instruction is organized around the science and engineering practices (National Research Council, 2012) and operationalized in the Next Generation Science Standards (NGSS Lead States, 2013). The framework’s eight practices range from asking questions, through planning and carrying out investigations and analyzing and interpreting data, to developing and using models, constructing explanations, and arguing from evidence. They provide a shared, standards-aligned structure onto which AI instruments can be mapped. Anchoring AI in scientific practices keeps the technology in the service of scientific inquiry: each AI instrument is placed at the practice it genuinely supports, and the practice itself, whether observation, analysis, or modeling, remains what the student is engaging in.

2.2 Operating Real Instruments

A defining feature of the approach is that students operate genuine AI instruments that are pedagogically bounded and aligned, e.g. via reduced complexity, rather than reading about or abstractly discussing AI in science. An AI instrument in this sense is something the student aims, calibrates, and must understand in its essentials, such as a clustering algorithm that the student runs and then critically evaluates. Operating the instrument constitutes authentic scientific practice; understanding how it behaves and where it fails constitutes discipline-based AI literacy. This understanding is primarily functional rather than mechanistic: mechanistic understanding concerns how an instrument works internally, whereas functional understanding concerns what it does, when its output can be trusted, and how it can fail (Hmelo-Silver & Pfeffer, 2004; Machamer et al., 2000). A consistent boundary runs through every instrument: it computes, but the student interprets. The instrument reports the structure it finds, yet what that structure means, and whether it can be trusted, is the student’s to decide. Unlike the microscope, whose workings a student can in principle inspect, AI instruments are often evaluated by their outputs without full mechanistic transparency. For a generative AI model, functional evaluation may be the primary form of evaluation available to the student, similar to how scientists often evaluate an opaque instrument through its outputs. A more transparent instrument such as clustering can in comparison be inspected more directly. Either way, the student judges the outcome the AI instrument produces.

The approach rests on two design commitments. First, each AI instrument is, while being a genuine scientific tool, pedagogically bounded: its interface and controls are simplified, while its core scientific function and the need for student interpretation are preserved. Second, each instrument is taught together with a characteristic reflection point—a consequential assumption, decision, or failure mode that prompts critical evaluation. Because an instrument works on the data it is given, reflection may concern the suitability of those data as well as the output itself. For clustering, for example, are the groups supported by the data or shaped by the variables included and the number of groups requested? For regression, does the selected function describe the data, or force a relationship through scattered points? For a generative model, did it produce what was specified, and is the output scientifically plausible? In the platform, this commitment is explicitly built into the AI instrument itself: an AI instrument should not be used for a scientific practice without first considering its potential reflection points. This is not a new form of critical thinking; it is the inherent skepticism of science, the same attention to controls, replication, and limits of traditional scientific instruments, directed at a new class of instruments. That continuity is what keeps the AI instrument within science education, with AI literacy embedded in the discipline, rather than constituting a separate unit on AI.

3 AI Instruments Across Scientific Inquiry

This article focuses on the investigative core of scientific inquiry: observing, analyzing, and modeling. For each of these, we describe one AI instrument in some detail and identify related instruments more briefly, noting how AI is used for that practice in current research.

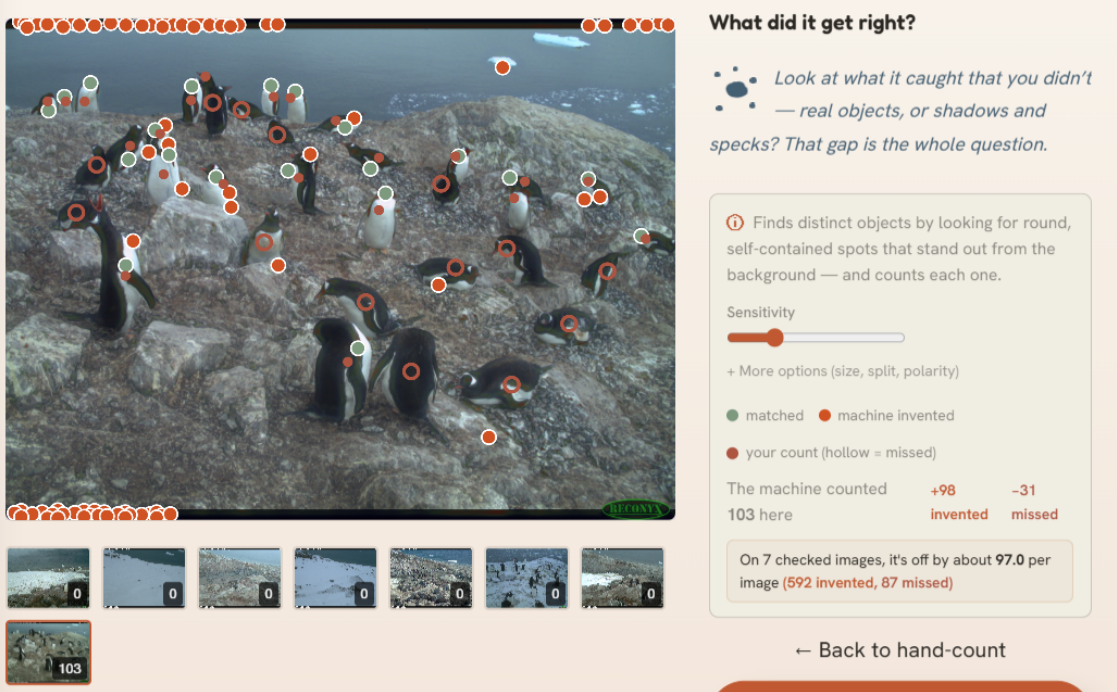
3.1 Observing: Turning the World into Data

In ecology, computer vision detects and counts animals in aerial and camera-trap imagery, a task that previously required years of manual effort. The computational methods range from classical computer-vision detectors that threshold an image and count the resulting blobs (Chabot & Francis, 2016) to deep neural networks that identify, count, and describe animals across millions of images (Norouzzadeh et al., 2018). Our first AI instrument is a pedagogically bounded version of this approach: students count objects across images, adjust the sensitivity, and compare the automated count with their hand count (Figure 1). The reflection point asks whether the instrument detected the intended objects and how characteristics of the images it was given, such as contrast, resolution, or framing, affected false detections and missed objects. A confident count can still be wrong. Related observing instruments extend this work beyond still images: object tracking follows individual animals across video frames (Romero-Ferrero et al., 2019), and audio classification identifies species from their calls (Kahl et al., 2021).

Table 1

Three Illustrative AI Instruments Within Scientific Inquiry, Each Anchored to an NGSS Practice, an Authentic Scientific Example, and an Exemplary Reflection Point

Scientific practice (NGSS)	AI instruments	Authentic scientific example	Exemplary reflection point
Planning and Carrying Out Investigations (focus: observing)	Computer vision: object detection & counting; object tracking; audio classification	Counting animals in camera-trap images (Norouzzadeh et al., 2018)	Did it count the intended objects, and how did characteristics of the images affect false detections and missed objects?
Analyzing & Interpreting Data	Clustering; regression	Finding structure in astronomical and genomic data (Jordan & Mitchell, 2015)	Are the groups supported by the data, or shaped by the variables included and the value of k ?
Developing & Using Models	Generative modeling; symbolic regression; ML surrogates	Designing novel proteins and molecules from a specification (Watson et al., 2023)	Did it build what was specified, and do scientific constraints and available evidence support the intended property?



What did it get right?

Look at what it caught that you didn't — real objects, or shadows and specks? That gap is the whole question.

① Finds distinct objects by looking for round, self-contained spots that stand out from the background — and counts each one.

Sensitivity

+ More options (size, split, polarity)

● matched ● machine invented
● your count (hollow = missed)

The machine counted +98 -31
103 here invented missed

On 7 checked images, it's off by about 97.0 per image (592 invented, 87 missed)

← Back to hand-count

Figure 1. Observing: Computer-vision counter

What it does: Detects and counts objects in an image; the student picks a detection method and tunes its sensitivity.

Student's task: Choose a method (edges, color, blob...), set the threshold, and compare the instrument's count with a hand count of the same image.

Built-in check: Overview of matched, falsely detected, and missed objects. The tool prompts students to consider how characteristics of the input image affected the count.

Across domains: The same approach can be used to count animals in wildlife photographs, cells in a micrograph, or craters in a surface image, although defining the target and interpreting the count remain domain-specific.

3.2 Analyzing and Interpreting Data

In data-rich fields such as astronomy and genomics, AI can identify candidate patterns in datasets too large to analyze by hand: it clusters data and fits relationships between measured quantities (Jordan & Mitchell, 2015; Wang et al., 2023). Our second AI instrument is a reduced version of clustering. Clustering sorts records into as many groups as the student asks for, by similarity; its reflection point is whether those groups were discovered or imposed by that choice, which the student can test by choosing a different number and evaluating mathematical and theoretical fit (Figure 2). A related instrument, regression, fits a function to the data and raises a parallel reflection point: whether a curve is being forced through scatter, and whether it can be trusted beyond the range of the data. The clustering instrument explicitly fosters reflection on the AI instrument: a clustering suggested by AI is not necessarily scientifically meaningful.

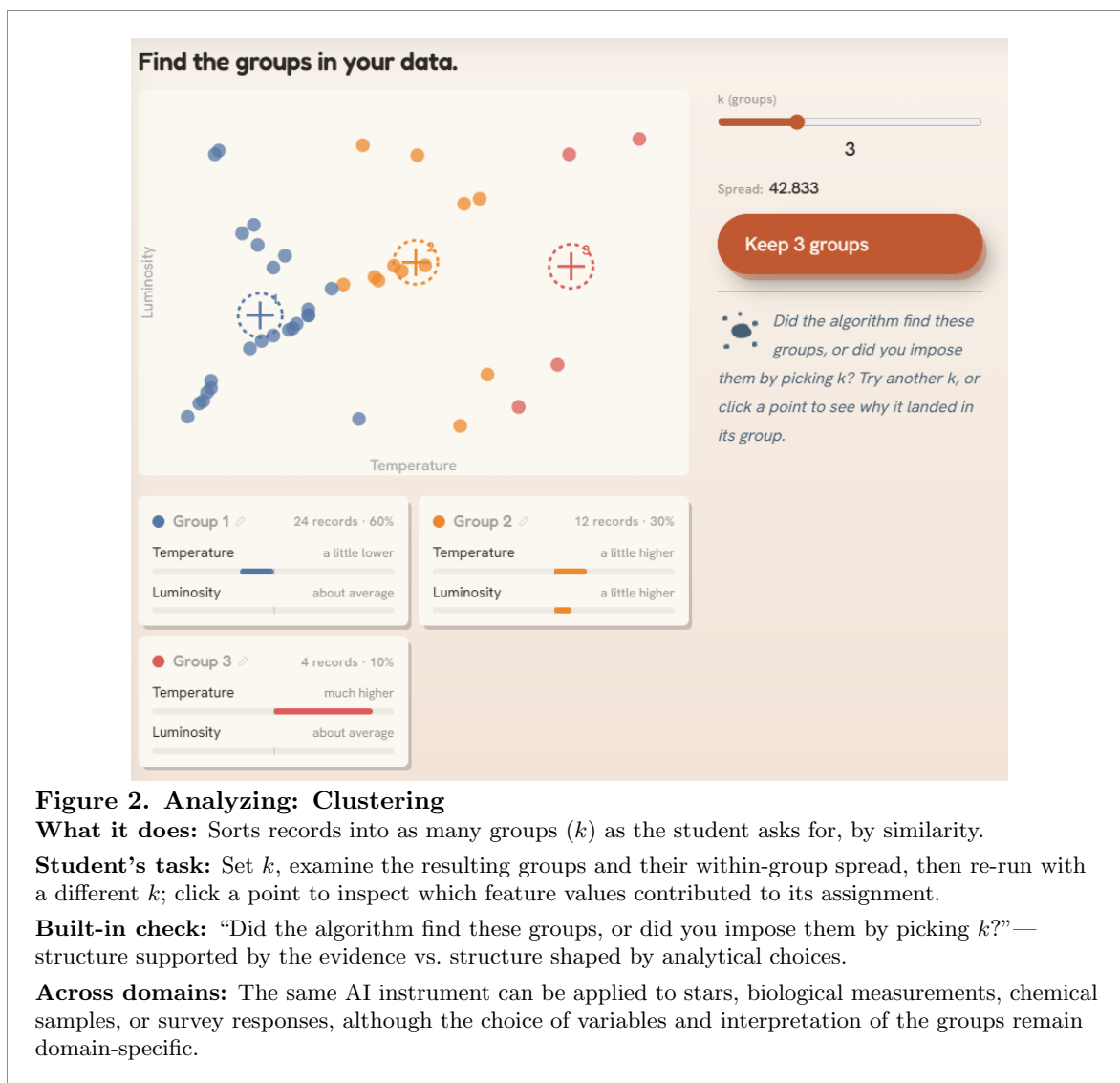


Figure 2. Analyzing: Clustering

What it does: Sorts records into as many groups (k) as the student asks for, by similarity.

Student’s task: Set k , examine the resulting groups and their within-group spread, then re-run with a different k ; click a point to inspect which feature values contributed to its assignment.

Built-in check: “Did the algorithm find these groups, or did you impose them by picking k ?”—structure supported by the evidence vs. structure shaped by analytical choices.

Across domains: The same AI instrument can be applied to stars, biological measurements, chemical samples, or survey responses, although the choice of variables and interpretation of the groups remain domain-specific.

3.3 Developing and Using Models

Generative models now design novel proteins and molecules from a target specification, producing candidate structures that need not occur in nature (Watson et al., 2023). Our third illustrative AI instrument is a pedagogically bounded version of generative design, set within the practice of developing and using models (Schwarz et al., 2009). The student can select or describe the features they think

are responsible for a property, such as which structural groups make a molecule taste sweet (Yun et al., 2026) or dissolve in water, and the AI instrument returns several candidate structures, each a generated structure to be evaluated. The generation itself is a “black box,” so the student’s task lies in evaluation and critique: weighing the candidates against the evidence. Its reflection point asks whether the instrument built what the student specified, whether the structure is scientifically plausible, and whether available evidence supports the intended property (Figure 3). This again is meant to foster the students’ explicit reflection on AI in the scientific inquiry. Two related instruments extend the modeling toolbox more briefly: symbolic regression, which has AI propose an equation rather than only a curve and raises the reflection point of whether the formula is a genuine law or an elaborate fit to noise (Schmidt & Lipson, 2009), and machine-learning surrogates, which give a fast approximation in place of an expensive simulation but raise the question of whether that approximation holds beyond its training range (Rasp et al., 2018). What the generative AI instrument adds is specific to this technology: aiming a model at an intended outcome and judging the result against that intent, not against how convincing it looks.

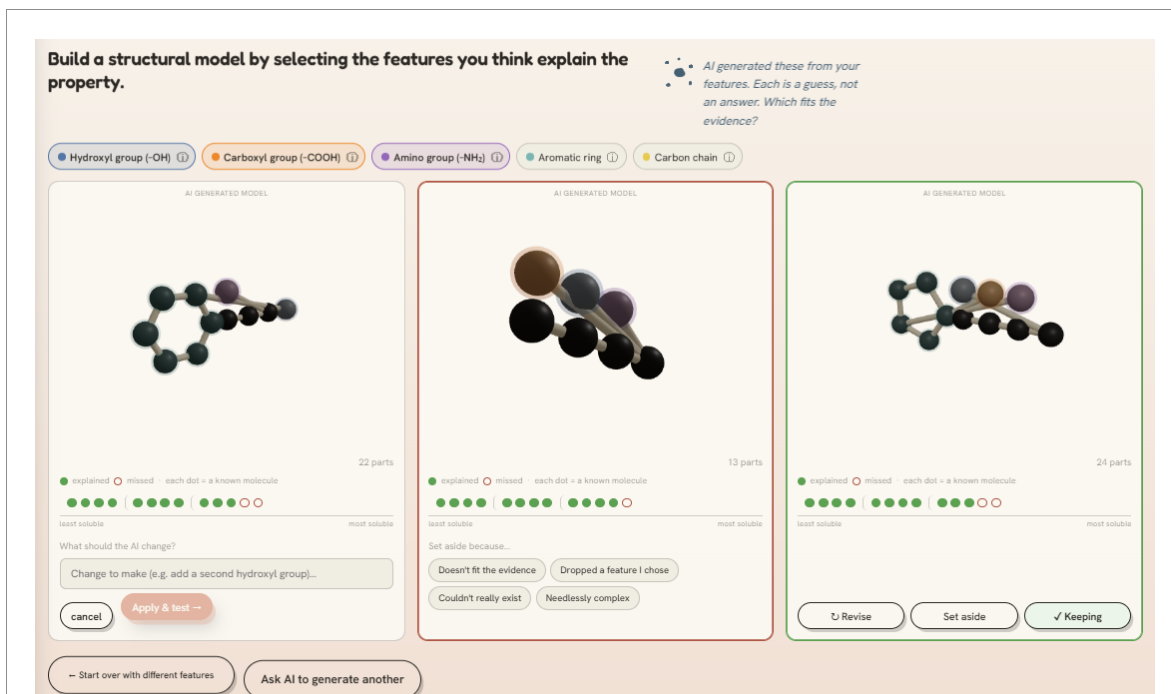


Figure 3. Modeling: Generative modeler

What it does: Turns a feature specification into candidate structures for the student to evaluate.

Student’s task: Choose the features thought to cause a property (e.g., solubility), generate candidates, and weigh each against the evidence—keep, revise, or set aside with a reason.

Built-in check: Did it build what you specified? Is the structure scientifically plausible, and does available evidence support the intended property? The “set aside because...” reasons make the relevant limitations explicit.

Across domains: The same AI instrument can propose candidate structures from a specification across different domains, although the constraints and criteria used to evaluate them remain domain-specific.

3.4 Asking Questions and Arguing from Evidence

Two practices often precede and follow the three illustrative functions described above, and AI has a genuine role in each. At the outset, in asking questions and planning investigations, AI can survey the literature, suggest candidate hypotheses, and propose experimental designs (Sourati & Evans, 2023; Wang et al., 2023). At the conclusion, in arguing from evidence and communicating results, it can compare findings across cases and assist in drafting figures and explanations. Each of these uses

carries its own reflection point, ranging from fabricated citations at the outset to a fluent summary that obscures a consequential disagreement at the end. We note them here for completeness, but the focus of this article remains the three illustrative instruments described above, where the role of AI is most concrete and science-specific.

3.5 Agentic AI as Collaborator

The three AI instruments above are illustrative examples of bounded tools operated by the student within a scientific practice. Increasingly, though, (partially) autonomous agents that follow a goal across many steps of scientific inquiry are becoming more common in the sciences (Gao et al., 2024). Such agentic AI can work across an entire inquiry process rather than within a single practice: it plans the steps, calls the individual instruments in turn, reads their results, and revises its approach. Such systems already design and run experiments (Boiko et al., 2023), propose and refine their own hypotheses (Google DeepMind, 2025), and carry out an investigation through its full cycle with little human intervention (Lu et al., 2024; Szymanski et al., 2023). What makes such agentic AI a collaborator rather than an instrument is not the range of practices it covers but the judgment it exercises within them. The researcher is left with setting the goal, directing, overseeing, and safeguarding the work, and answering for what the results mean. These systems lie beyond the single-practice AI instruments described above, and we treat agentic AI here as a horizon that instruction must prepare for.

As agentic AI becomes part of how science is practiced, students need to understand that it exists and how it works to gain a comprehensive scientific literacy (Nehm & Kubsch, 2026). Yet the autonomy that makes it useful is also what makes early reliance risky: an agentic system that can carry out a whole inquiry invites the student to hand it over completely. As it takes over large portions of the inquiry process, the student’s competence might decay or never properly develop through disuse. But that competence, core scientific literacy, is exactly what is needed to supervise the system, notice its errors, and step in when it fails (Bainbridge, 1983; Parasuraman & Manzey, 2010).

The order of instruction, therefore, might matter. To judge what agentic AI produces and whether it can be trusted, a student needs some scientific literacy and DAIL (Zhai, 2025). A scientific inquiry an AI agent carries out autonomously is one that the student might have little chance to learn from. Scientific literacy can grow alongside a student’s use of AI instruments, but the capacity to evaluate an agent’s outputs should develop before substantial parts of an investigation are handed over. Agentic AI therefore does not replace direct observation through a microscope, nor the inquiry carried out with the individual AI instruments of Sections 3.1 to 3.4; it builds on them. Relied upon before sufficient scientific literacy and DAIL are in place, agentic AI risks dependence instead of understanding, leaving students who can neither judge what the AI agent returns nor acquire the scientific literacy the instruction exists to develop.

4 Discussion

4.1 Educational Opportunities

An argument for using AI instruments in science education is authenticity: pedagogically bounded instruments can preserve genuine scientific practices, such as counting populations from images, clustering complex data, and generating candidate models, while leaving interpretation and evaluation to the student. Because the AI instruments are situated within NGSS practices, the AI content does not compete with the science curriculum but extends it. Equally important, the approach treats the limitations and evaluation of the AI as an integral aspect of the instruction rather than as a hidden liability. Each AI instrument is paired with a reflection about how it can be wrong, making reasoning about its flaws an explicit part of the work, which is the core of scientific inquiry.

4.2 Implications for AI Literacy

This design enacts discipline-based AI literacy in the sense recent work has called for (Knoth et al., 2024). Rather than presenting AI concepts like “clustering” in the abstract or focusing on the computational mechanics, and relying on transfer, the instruction situates each concept within an authentic AI instrument: students learn about generative modeling by specifying and evaluating generated candidates, and they learn its risks by seeing what happens when the features they choose fail to account

for the evidence. The approach complements rather than replaces general frameworks such as those of Long and Magerko (2020) and Touretzky et al. (2019). What accumulates across units is not a catalog of algorithms but an understanding of how AI changes the practice of science, a form of literacy that is difficult to obtain from a standalone AI literacy course but necessary for a scientifically literate public.

4.3 Human-AI Collaboration

When students use the individual AI instruments, the division of work is clear. Each instrument supports a distinct scientific practice. The student sets the intent and decides whether to trust what comes back. Agentic AI, by contrast, has the capability to decide what to run next, when a result is good enough, and which line of inquiry to drop. Those are the student’s decisions when the AI instruments are used by the student. With an AI agent in the loop, the line between the student’s thinking and the system’s “reasoning” becomes increasingly blurred.

Once the AI agent becomes a partly independent collaborator, validation becomes more difficult. A student can test a single clustering directly against the data, but validating a whole inquiry built from the agent’s chain of choices is more demanding because each decision, intermediate output, and dependency must also be inspected. The student may come to trust the agent rather than reflect on and check it. This is where we see the contribution of the AI instruments (Sections 3.1–3.3). Operating them is how a student learns the reflection and critical evaluation that an AI agent will demand, both to direct it and to question its outputs. Whether that critical evaluation carries over from a single set of AI instruments to a full inquiry remains to be understood. This reflects a careful, accountable partnership that practicing scientists are themselves still working out (Wang et al., 2023).

4.4 Challenges and Limitations

Among the challenges this vision raises, the first is over-reliance. AI instruments are fluent enough that students may take their output at face value and skip the reasoning the lesson is intended to develop. The explicit focus on reflection points is a direct response, but it is effective only insofar as the surrounding instruction prompts students to apply it. The accompanying assistant is constrained for the same reason: its responses are tied to the activity materials and the student’s own work, and it redirects requests for the answer back to the evidence rather than supplying it. The reliability of each instrument also depends on the data it is given, making the selection of suitable images, variables, examples, or constraints an important part of instructional design. The second challenge is teacher preparation. Designing lessons that place AI instruments appropriately within scientific inquiry and guiding the reflection each instrument is meant to spark demands considerable expertise at a time when many teachers are still developing their own AI literacy. A third is evidential: the approach is so far a design argument, and whether operating these instruments produces the scientific and AI literacy we claim remains to be tested in classrooms. These challenges are not disqualifying, but they mark the conditions under which careful design and further research are needed.

5 Conclusion

AI has become a new generation of scientific instruments, distributed throughout scientific inquiry. A science education that ignores this development leaves students unprepared for science as it is now practiced. The approach we describe anchors each AI instrument to an NGSS scientific practice and pairs it with an explicit way to reflect and critically evaluate how it can mislead. It follows two aims at once: authentic scientific inquiry and DAIL that develops into a working map of where and how AI enters science. The vision is concrete because its AI instruments do real scientific work, and it is sustainable because the skepticism it cultivates is the established skepticism of science, now directed at a new class of instruments. Beyond these instruments, researchers already work with agentic AI that runs across the whole scientific inquiry. Students should understand this emerging reconfiguration of scientific inquiry and how to reflect on the AI agent’s role (Nehm & Kubsch, 2026). As argued in Section 3.5, the order of instruction matters: students should develop foundations in scientific inquiry and its AI instruments before relying on increasingly autonomous AI agents, or they risk handing the science to a system they can neither adequately check nor take over when it fails.

Declarations

CRedit authorship contribution statement

Arne Bowersdorff: Conceptualization, Software, Writing—original draft, Writing—review & editing, Visualization.

Matias Rojas: Writing—review & editing.

Xiaoming Zhai: Conceptualization, Supervision, Funding acquisition, Writing—review & editing.

Competing interests

The authors declare no competing interests.

Funding

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305C240010 (PI Zhai). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Acknowledgements

We want to thank M. Yun, K. J. Crippen, C. Xie, & D. Bulseco for their valuable insights and pioneering work on the GenAI molecule modelling approach.

Use of AI in the writing process

During the preparation of this work, the authors used ChatGPT and Anthropic Claude in order to improve the readability and language of sentences as authors are not native English speakers. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

- Bainbridge, L. (1983). Ironies of automation. *Automatica*, 19(6), 775–779. [https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/10.1016/0005-1098(83)90046-8)
- Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. *Nature*, 624(7992), 570–578. <https://doi.org/10.1038/s41586-023-06792-0>
- Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3), 199–231. <https://doi.org/10.1214/ss/1009213726>
- Chabot, D., & Francis, C. M. (2016). Computer-automated bird detection and counts in high-resolution aerial images: A review. *Journal of Field Ornithology*, 87(4), 343–359. <https://doi.org/10.1111/jofo.12171>
- Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A., Schwarz, J. R., Ektefaie, Y., Kondic, J., & Zitnik, M. (2024). Empowering biomedical discovery with AI agents. *Cell*, 187(22), 6125–6151. <https://doi.org/10.1016/j.cell.2024.09.022>
- Google DeepMind. (2025). *Towards an AI co-scientist*. arXiv. <https://arxiv.org/abs/2502.18864>
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- Herdlika, A., & Zhai, X. (2024). Artificial intelligence-based scientific inquiry. In X. Zhai & J. Krajcik (Eds.), *Uses of artificial intelligence in STEM education* (pp. 179–197). Oxford University Press. <https://doi.org/10.1093/oso/9780198882077.003.0009>
- Hmelo-Silver, C. E., & Pfeffer, M. G. (2004). Comparing expert and novice understanding of a complex system from the perspective of structures, behaviors, and functions. *Cognitive Science*, 28(1), 127–138. https://doi.org/10.1207/s15516709cog2801_7
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. <https://doi.org/10.1126/science.aaa8415>

- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Kahl, S., Wood, C. M., Eibl, M., & Klinck, H. (2021). BirdNET: A deep learning solution for avian diversity monitoring. *Ecological Informatics*, 61, 101236. <https://doi.org/10.1016/j.ecoinf.2021.101236>
- Knoth, N., Decker, M., Laupichler, M. C., Pinski, M., Buchholtz, N., Bata, K., & Schultz, B. (2024). Developing a holistic AI literacy assessment matrix: Bridging generic, domain-specific, and ethical competencies. *Computers and Education Open*, 6, 100177. <https://doi.org/10.1016/j.caeo.2024.100177>
- Long, D., & Magerko, B. (2020). What is AI literacy? competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. <https://doi.org/10.1145/3313831.3376727>
- Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). *The AI scientist: Towards fully automated open-ended scientific discovery*. arXiv. <https://arxiv.org/abs/2408.06292>
- Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1), 1–25. <https://doi.org/10.1086/392759>
- National Research Council. (2012). *A framework for K–12 science education: Practices, crosscutting concepts, and core ideas*. The National Academies Press. <https://doi.org/10.17226/13165>
- Nehm, R. H., & Kubsch, M. (2026). AI in contemporary scientific practice: Implications for AI-integrated science education. In X. Zhai & K. J. Crippen (Eds.), *Advancing AI in science education: Envisioning responsible and ethical practice* (pp. 27–48, Vol. 3). Springer. https://doi.org/10.1007/978-3-032-16871-9_2
- NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academies Press. <https://doi.org/10.17226/18290>
- Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*, 115(25), E5716–E5725. <https://doi.org/10.1073/pnas.1719367115>
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381–410. <https://doi.org/10.1177/0018720810376055>
- Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115(39), 9684–9689. <https://doi.org/10.1073/pnas.1810286115>
- Romero-Ferrero, F., Bergomi, M. G., Hinz, R. C., Heras, F. J. H., & de Polavieja, G. G. (2019). Idtracker.ai: Tracking all individuals in small or large collectives of unmarked animals. *Nature Methods*, 16(2), 179–182. <https://doi.org/10.1038/s41592-018-0295-5>
- Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
- Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science*, 324(5923), 81–85. <https://doi.org/10.1126/science.1165893>
- Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Acher, A., Fortus, D., Shwartz, Y., Hug, B., & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching*, 46(6), 632–654. <https://doi.org/10.1002/tea.20311>
- Sourati, J., & Evans, J. A. (2023). Accelerating science with human-aware artificial intelligence. *Nature Human Behaviour*, 7(10), 1682–1696. <https://doi.org/10.1038/s41562-023-01648-z>
- Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K., Zeng, Y., & Ceder, G. (2023). An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990), 86–91. <https://doi.org/10.1038/s41586-023-06734-w>
- Touretzky, D., Gardner-McCune, C., Martin, F., & Seehorn, D. (2019). Envisioning AI for K–12: What should every child know about AI? *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), 9795–9799. <https://doi.org/10.1609/aaai.v33i01.33019795>

- Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., Anandkumar, A., Bergen, K., Gomes, C. P., Ho, S., Kohli, P., Lasenby, J., Leskovec, J., Liu, T.-Y., Manrai, A., & Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972), 47–60. <https://doi.org/10.1038/s41586-023-06221-2>
- Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., Ahern, W., Borst, A. J., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Hanikel, N., Pellock, S. J., Courbet, A., Sheffler, W., Wang, J., Venkatesh, P., Sappington, I. D., Vázquez Torres, S., & Baker, D. (2023). De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976), 1089–1100. <https://doi.org/10.1038/s41586-023-06415-8>
- Yun, M., Crippen, K. J., Xie, C., & Bulseco, D. (2026). From prompt engineering to modeling: Secondary science teachers’ use of generative AI to engineer sweetener molecules. *Chemistry Education Research and Practice*. <https://doi.org/10.1039/D6RP00064A>
- Zhai, X. (2025). *DAIL: Discipline-based artificial intelligence literacy*. SSRN. <https://doi.org/10.2139/ssrn.5745703>