Accelerating Scientific Discovery With AI-Aided Automation

Tapio Schneider, *California Institute of Technology, Pasadena, CA, 91125, USA*
Ilkay Altintas, *University of California, San Diego, La Jolla, CA, 92093, USA*
Daniel Atkins, *University of Michigan, Ann Arbor, MI, 48109, USA*

AI-aided design of experiments and observations, together with robotic instrumentation and automated learning from data, has the potential to transform science and propel it forward with unprecedented speed.

In 1620, Francis Bacon urged scientists not only to observe nature, but also to actively manipulate it to uncover its secrets.[1] This foundational principle of empirical science has guided centuries of progress, forming a virtuous circle of iterative exploration: an explicit or implicit model of a system is used as the basis for designing experiments or observations; the resulting data are harnessed to improve the model through calibration, revision, or even a complete overhaul; and the cycle repeats (Figure 1). In this context, a "model" can refer to a specific instance of a theory, such as a model based on the general theory of relativity to explain gravitational waves from black hole collisions. It can also encompass empirical models, such as an empirical model of disease progression, or a combination of both, such as a climate model that incorporates physical laws alongside empirical closure relations for small-scale processes. The term "data" encompasses information obtained through observations, simulation studies, or laboratory experiments. The process of knowledge generation within this loop can begin at any stage, whether it be the development of a new model prompting the generation of new data or the acquisition of new data inspiring the design or modification of an existing model.[2]

Over the past few decades, computing has increasingly assumed a crucial role within this knowledge generation loop. It fulfills a wide range of functions, including data processing and analysis, experimental design and control, and data generation through simulations. This expansion of computing's involvement has led to the emergence of entirely new disciplines, such as computational biology and computational astrophysics. Despite these advancements, human interventions have remained a vital component of the loop, which has inevitably limited the rate of iteration.

We are currently in a transformative phase, facilitated by advancements in AI, computing, and the automation of laboratory and observing systems, as outlined in a recent consensus study by the National Academies of Science, Engineering, and Medicine (NASEM).[2] The simultaneous progress in AI, computing, and automation has the potential to remove human intervention within the loop, automating and accelerating the rate of iteration through the knowledge generation loop, often by orders of magnitude. AI, broadly understood to include tools ranging from Bayesian learning to deep learning, now enables us not only to learn about parameters and parametric functions within models but even to derive mathematical theorems[3]; discover the governing equations of models[4,5]; or generate "equation-free" models, such as AlphaFold, which predicts the 3-D structure of a protein from its amino acid sequence.[6] AI can also be used for "active learning," that is, to design experiments or observations that maximize information gain on uncertain aspects of a model. This can be achieved through solutions to Bayesian optimal design problems, exploration algorithms as used in reinforcement learning, or generative AI models.

In addition to AI advancements, automation and robotic instrumentation are transforming laboratories and observational devices. Telescopes, for instance, are now routinely controlled remotely by computers,

This article was invited by the 2023 Editor-in-Chief, Lorena Barba.

1521-9615 © 2024 IEEE
Digital Object Identifier 10.1109/MCSE.2024.3352451
Date of current version 12 April 2024.

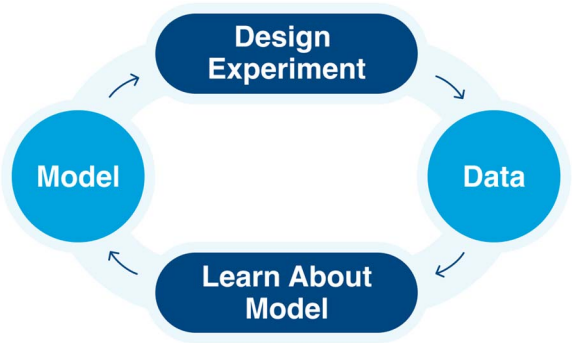

FIGURE 1. Science has progressed by iteration through a virtuous circle: models inform the design of experiments or observations; the data they produce are harnessed to inform, update, or revise models. In many fields, AI tools together with automated data acquisition now enable automation and acceleration of the entire loop, often by orders of magnitude.

enabling automated target selection. Similarly, biological and chemical laboratories employ microfluidic devices, facilitating automated experiments at a higher throughput than manual methods allow. Consequently, in areas critical to human welfare, such as drug discovery for combating infectious diseases or the development of climate models for predicting our future, the potential for rapid and transformative scientific and engineering progress is within reach.

AI-AIDED AUTOMATION IN ACTION

Consider a few pioneering examples:

- *Bioengineering*: Directed evolution is an optimization process for protein engineering that mimics and expedites natural evolution in vitro. Recognized with the Nobel Prize in Chemistry in 2018, it involves an interactive procedure starting from a protein state, exploring random mutations, selecting desirable outcomes through screening or selection, and repeating until a protein with the desired properties is obtained. This has sped up evolution from timescales of millions of years to weeks. AI is now further accelerating directed evolution: machine learning models trained on tested variants predict good candidates for further exploration, significantly reducing the experimental burden.[7] The entire process, including screening and selection, can be automated, providing a fast and efficient means to evolve proteins for technological, scientific, and medical applications.[8]
- *Chemistry and materials science*: The discovery process for new medicines, agrochemicals, and everyday materials involves conceiving reactions or materials, their subsequent synthesis, and their testing or characterization in the laboratory. This process so far has relied on human exploration and experimentation. AI can now speed up the exploration and design process, enabling autonomous high-throughput synthesis[9] and rapid characterization. With the aid of robotic machinery and self-driving laboratories, we can now accelerate the pace of discovery and improve the quality of the molecules and materials of tomorrow.[10]
- *Astronomy*: Astronomy, being inherently data intensive, has already witnessed AI-driven discoveries in datasets from sky surveys across various wavelength bands. Additionally, many telescopes are now operated robotically and can be controlled remotely. Consequently, astronomy is poised to take the next step, where AI will make adaptive choices about survey strategies, such as target selection, based on maximizing information gain about the underlying models.[11]

It is evident that sweeping changes are underway—if not in the entire scientific enterprise, then at least within specific scientific domains. This evolution promises a leap in scientific productivity, particularly in areas where the predictive quality of models and the added value of new data can be quantified, and both can be iteratively optimized by cycling through an automated knowledge discovery loop. The move toward automated, AI-driven scientific processes brings with it improved reproducibility, replicability, and shareability that accompanies workflows driven by code.[12] This, in turn, can focus the attention of scientific communities and accelerate information flow within them, signaling a new paradigm in how we approach, understand, and harness the power of scientific exploration.

REALIZING THE POTENTIAL OF AI-AIDED AUTOMATION

To fully harness the potential of AI-aided automation in science and engineering, deliberate planning and the realignment of incentives are necessary. While models and data generation methods have strong domain-specific components (the nodes of the graph in Figure 1), the approaches to designing experiments and observations and to learning from data transcend scientific fields (the edges of the graph in Figure 1), similar to the way least squares fitting of data can be useful in any domain. For AI-aided automation to have widespread

impact across disciplines, it is essential to make AI methods for science and engineering accessible and promote their extensive use, fostering methodological convergence within and among scientific fields.

SIMILAR TO HOW CALCULUS AND STATISTICS ARE FUNDAMENTAL COMPONENTS OF THEIR EDUCATION, LEARNING ABOUT AI PRINCIPLES AND APPLICATIONS SHOULD BE EQUALLY EMPHASIZED.

The availability of open source AI libraries, such as TensorFlow and PyTorch, has resulted in a surge in AI usage, including in science and engineering. However, it is important to note that these off-the-shelf AI tools may not always be suitable for scientific and engineering purposes. For instance, they primarily focus on supervised learning, which relies on labeled input–output pairs of a process to train a model for it. However, in many science and engineering problems, obtaining such labeled data at relevant scales or in sufficient quantities is challenging. We require AI tools that are implemented in professionally developed and well-maintained open source software packages, specifically tailored for science and engineering applications, where data typically are noisy, are heterogeneous, and have missing values. These tools should align with the needs of the scientific community, similar to the widely adopted packages in the machine learning community.

The NASEM report on Automated Research Workflows[2] put forward several recommendations that, if put into practice, would expedite the achievement of the transformative potential of AI-aided automation in science and engineering. These recommendations include the following:

1) *Train early-career researchers in AI tools*: The next generation of scientists and engineers needs to be fluent in methods of AI and computing. AI and computing are the new calculus. Similar to how calculus and statistics are fundamental components of their education, learning about AI principles and applications should be equally emphasized. Notebook platforms such as Jupyter and Pluto.jl provide accessible entry points for learning about AI methods and designing automated research workflows.
2) *Develop cyberinfrastructure for AI*: To enable the widespread adoption of AI tools in science and engineering, the availability of user-friendly and professionally maintained software is crucial. Currently, the focus on peer-reviewed publications hinders the development of such software. To address this, sustained funding should be allocated for research software engineers who work collaboratively with science and engineering teams, prioritizing the development of high-quality open source software. Funders and research institutions can play a role by making open source software a requirement in projects and providing the necessary funding.
3) *Incentivize team science*: Research as a public good continues to attract talented early-career scientists and engineers, who enjoy working on important problems in cross-functional and cross-disciplinary teams. However, existing reward structures that prioritize publications and lead authorship, coupled with the limited number of principal investigator positions relative to the number of those with Ph.D. degrees, create challenges for careers in science and engineering outside the private sector. Stable career prospects should be provided for researchers who contribute to team success rather than solely focusing on individual publication output. This will facilitate the realization of the potential of AI-aided automation in science and engineering.
4) *Democratize data and facilitate access*: Accelerated adoption of AI-aided automation of science depends on equitable and sustainable access to findable, accessible, interoperable, and reusable (FAIR[a]) data. Breakthroughs often arise from unexpected correlations, potentially among diverse datasets, and the discovery process is greatly enhanced by FAIR data. With scientific data volumes increasing exponentially, traditional methods of downloading and local processing become impractical. Therefore, data should be provided in cloud-optimized formats to enable processing directly where the data are stored, in the cloud. Funders, research institutions, and publishers can incentivize FAIR and cloud-optimized data by implementing open research requirements and providing funding to support these initiatives.

We stand at the cusp of the next scientific revolution, where the frictional cost of human manipulations in iterating through the virtuous cycle of scientific

[a]https://www.go-fair.org/fair-principles/

knowledge discovery can vanish, and the entire loop can be automated and accelerated. Guiding the transformation will require deliberate nurturing by funders and research institutions. The structures necessary to unlock its full potential do not neatly align with the traditional paradigm of the individual principal investigator that has prevailed for centuries. The advent of AI-aided automation in science also raises thought-provoking questions about the agency and creativity of scientists as well as the role of serendipity within automated research workflows. In this new landscape, scientists will serve as architects of strategy and instrumentation, designing systems that execute and iterate automatically, harnessing the power of AI and automation to exponentially accelerate the success model of science that has served us remarkably well for the past 400 years.

ACKNOWLEDGMENTS

We thank the members of the National Academies study on Automated Research Workflows for their contributions to the report on which this perspective is based, Frances Arnold for her insights on directed evolution and its automation, and Sarah Reisman for helpful discussions on automated chemical synthesis. The National Academy of Science study on Automated Research Workflows was generously supported by Eric and Wendy Schmidt by recommendation of Schmidt Futures.

REFERENCES

1. I. Hacking, *Representing and Intervening*. Cambridge, U.K.: Cambridge Univ. Press, 1983.
2. National Academies of Sciences, Engineering, and Medicine. *Automated Research Workflows for Accelerated Discovery: Closing the Knowledge Discovery Loop*. Washington, DC, USA: The National Academies Press, 2022.
3. A. Davies et al., "Advancing mathematics by guiding human intuition with AI," *Nature*, vol. 600, no. 7887, pp. 70–74, 2021, doi: 10.1038/s41586-021-04086-x.
4. S. H. Rudy et al., "Data-driven discovery of partial differential equations," *Sci. Adv.*, vol. 3, no. 4, 2017, Art. no. e1602614, doi: 10.1126/sciadv.1602614.
5. M. Raissi and G. E. Karniadakis, "Hidden physics models: Machine learning of nonlinear partial differential equations," *J. Comput. Phys*, vol. 357, pp. 125–141, Mar. 2018, doi: 10.1016/j.jcp.2017.11.039.
6. J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021, doi: 10.1038/s41586-021-03819-2.
7. K. K. Yang et al., "Machine-learning-guided directed evolution for protein engineering," *Nature Methods*, vol. 16, no. 8, pp. 687–694, 2019, doi: 10.1038/s41592-019-0496-6.
8. D. C. Miller et al., "Combining chemistry and protein engineering for new-to-nature biocatalysis," *Nature Synthesis*, vol. 1, no. 1, pp. 18–23, 2022, doi: 10.1038/s44160-021-00008-x.
9. Y. Shen et al., "Automation and computer-assisted planning for chemical synthesis," *Nature Rev. Methods Primers*, vol. 1, no. 1, p. 23, 2021, doi: 10.1038/s43586-021-00022-5.
10. D. P. Tabor et al., "Accelerating the discovery of materials for clean energy in the era of smart automation," *Nature Rev. Mater.*, vol. 3, no. 5, pp. 5–20, 2018, doi: 10.1038/s41578-018-0005-z.
11. A. Szalay, "The era of surveys and the fifth paradigm of science," in *Proc. AAS Meeting*, Washington, DC, USA: American Astronomical Society, 2019, p. 400.
12. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science*. Washington, DC, USA: The National Academies Press, 2019.

TAPIO SCHNEIDER is with the California Institute of Technology, Pasadena, CA, 91125, USA. Contact him at tapio@caltech.edu.

ILKAY ALTINTAS is with the University of California, San Diego, La Jolla, CA, 92093, USA. Contact her at ialtintas@ucsd.edu.

DANIEL ATKINS is with the University of Michigan, Ann Arbor, MI, 48109, USA. Contact him at atkins@umich.edu.