The Discovery Engine

Jessica Rumbelow*, Jugal Patel, Robbie McCorkell, Zohreh Shams,
Andrew Cusick, Arush Tagade, Leo Mckee-Reid, and Jack Foxabbott

Leap Laboratories

June 18, 2025

Abstract

We introduce the Discovery Engine, a general purpose automated system for scientific discovery. It combines deep learning with state-of-the-art interpretability techniques to identify complex, non-linear relationships in arbitrary datasets. This technology enables a shift from hypothesis-driven to data-driven discovery, and massively accelerates scientific enquiry by allowing a full exploration of the space of possible insights, free of bias and assumptions. It is hundreds of times faster than manual analysis, domain-agnostic, and data efficient – requiring only hundreds (rather than hundreds of thousands) of samples, and thereby making AI for science accessible in domains where data is limited or costly to obtain. In this paper we describe the Discovery Engine system and contrast it with contemporary approaches to AI for general scientific discovery (particularly LLM-driven methods, which form the bulk of alternative solutions).

1 Introduction

The scientific method has served as the foundation of empirical enquiry for over four centuries. However, increasing evidence suggests that this traditional paradigm faces unprecedented challenges. Despite exponential growth in research funding and personnel, the rate of transformative discoveries has decreased [3, 29, 7]. Concurrently, the replication crisis has revealed that a significant proportion of published findings are likely false [14, 27, 1]. Aside from underlining the ongoing problem of perverse incentives in academic publishing, these systemic issues suggest a need for fundamental methodological change.

This paper endorses a paradigm shift from hypothesis-driven to data-driven discovery, enabled by recent advances in artificial intelligence. We present our Discovery Engine, an automated system that leverages Leap Laboratories' interpretability research to automatically identify patterns in complex datasets without requiring predetermined hypotheses. This approach addresses key limitations of traditional (slow, biased) scientific methodology.

*Corresponding author: jessica@leap-labs.com

2 Limitations of the Current Scientific Paradigm

The traditional scientific method relies on researchers to formulate testable hypotheses based on existing knowledge and intuition. This approach is inherently limited by human cognitive capacity and subject to various biases. Researchers can explore only a tiny fraction of the potential hypothesis space, and this exploration is necessarily path-dependent, influenced by personal experience, training, and familiarity with existing literature [40, 31].

Furthermore, the reliability of hypothesis generation is compromised by the high rate of non-replicable findings in the scientific literature. When researchers base new hypotheses on flawed prior work, errors propagate through the scientific knowledge base. Large Language Models (LLMs) trained on scientific literature inherit and potentially amplify these issues, as they lack mechanisms for distinguishing valid from invalid findings [18, 36].

In general, traditional experimental design focuses on testing specific hypotheses, which constrains both the variables measured and the analytical approaches employed. This targeted approach, while efficient for testing predetermined ideas, severely limits the potential for serendipitous discovery. Confirmation bias further exacerbates this issue, as researchers may consciously or unconsciously design experiments and interpret results in ways that support their hypotheses [38, 28].

Finally, the 'publish or perish' paradigm in academia creates perverse incentives that prioritize quantity over quality, leading to p-hacking, selective reporting, and publication bias [26, 10, 8]. While addressing these institutional issues falls outside the scope of this work, their impact on scientific progress must be acknowledged as it directly contributes to the unreliability of scientific literature, which impacts many approaches to AI in the sciences.

3 Large Language Models for Scientific Discovery

As much of AI-for-science research outside of our own lab seems to focus on the application of LLMs (and agents built atop them), we include here a brief discussion of how we view these impressive models and their limitations when it comes to accelerating scientific discovery – and ultimately why we find them ill-suited for the job.

Direct, domain-specific applications of machine learning to specific scientific tasks (e.g. protein folding [15]) have proven extremely useful, but solve a different problem – one of automation to increase scientific *capability*, rather than the general problem of obtaining novel insight. For that reason they are not discussed in this paper, as we are primarily concerned with general purpose technologies for automating scientific discovery, rather than specific solutions in any one domain. Additionally, these kinds of foundation models are typically extremely data-hungry [44], and require rare expertise and financial resources to train, rendering this approach to scientific progress inaccessible to most scientists.

3.1 Current Capabilities and Applications

Large Language Models represent a transformative technology built on transformer architectures and trained to predict subsequent tokens across vast text corpora. Through reinforcement learning techniques, these models have been adapted to perform diverse tasks with remarkable proficiency. In the context of scientific research, LLMs demonstrate particular strength in tasks that fall *within* their training distribution. These include: generation of boilerplate code and analytical scripts; data visualization and plotting; organization and structuring of research notes; LaTeX formatting and document preparation; text editing and refinement; initial drafting of manuscripts; and explanation of well-established scientific concepts. Such capabilities have made LLMs invaluable tools for automating routine aspects of scientific work, potentially enhancing researcher productivity [19, 25].

However, despite many impressive capabilities, LLMs face critical limitations when applied to scientific discovery. They excel at producing text that appears credible and well-reasoned, regardless of factual accuracy. This capability, while useful for creative tasks, is risky in scientific contexts.

3.2 Fundamental Limitations of LLMs for Frontier Research

While careful prompt engineering can reduce hallucination rates, researchers must continuously verify LLM-generated content, including citations and factual claims, creating a demanding cognitive burden that undermines efficiency gains. At the frontier of research, where ground truth may be unknown or contested, distinguishing between accurate insights and convincing hallucinations becomes exceptionally challenging or impossible. We see this occur often in automated hypothesis generation systems, which produce numerous candidates – some of which prove useful, but many of which are ultimately misguided [5, 39, 12, 35].

Hallucination aside, LLMs inherit the biases and errors present in their training data [43]. Given the replication crisis affecting numerous scientific fields, models trained on published literature inevitably encode false findings and flawed methodologies. This creates a compound problem: even when LLMs accurately reproduce information from their training data, that information may itself be incorrect. This creates a dangerous feedback loop where each iteration potentially degrades rather than improves scientific quality. The prospect of academic repositories flooded with convincing but fundamentally flawed papers poses significant risks to scientific integrity. Review models trained on contradictory peer reviews introduce additional noise rather than clarity.

Finally, much is lost in translation: academic papers in empirical fields are themselves an extremely lossy abstraction over the underlying data, in which we hope to find meaningful insight.

The key insight here is that models trained on human-generated scientific text inevitably inherit the limitations, biases, and errors present in that text. Without mechanisms to transcend these inherited flaws, such systems risk accelerating the production of inherently flawed science rather than advancing genuine discovery. LLMs employed in

this way may well prove successful at automating the existing scientific status quo: a paradigm that has given us much, but so much less than it could. Scientific discovery is slowing and the reproducibility crisis is compounding. To fully realise the potential of artificial intelligence for scientific discovery, we need a new paradigm – one that we believe the Discovery Engine enables.

3.3 Strategic Use of LLMs in the Discovery Engine

Despite these concerns, we believe that language models do have their place in scientific discovery. We use LLMs exclusively to contextualize empirically-derived findings with respect to existing literature, an approach that takes advantage of their spectacular ability to synthesise textual information, while avoiding reliance on them for primary discovery, where they are least reliable. Because all of our primary discoveries emerge from direct analysis of empirical data rather than textual descriptions, findings are grounded in actual observation rather than linguistic patterns. We recognise that, for all its flaws, existing scientific literature does encode one thing faithfully: what humans already know, and what we care about knowing.

4 Data-Driven Discovery

Instead of automating the existing paradigm, we endorse a fundamental reorientation of the scientific process: rather than beginning with hypotheses, researchers should collect comprehensive data about phenomena of interest, and systematically analyse this data to reveal patterns.

Extracting novel insights from data is not a new idea – prior to the advent of big data, Knowledge Discovery in Databases (KDD), defined as 'the non-trivial extraction of implicit, previously unknown, and potentially useful information from data', was first discussed by Frawley et al [11] in 1991. KDD typically employs a number of statistical methods along with some more inherently interpretable machine learning techniques, such as decision trees, dimensionality reduction, and clustering. These methods have been effectively used for discovery, such as identifying unsuspected adverse drug reactions [2]. However, these methods often assume feature independence, cannot capture complex non-linear patterns, require manual feature engineering, and do not handle multimodal data well, limiting their potential for truly novel discovery on today's rich, complex datasets.

In contrast, deep neural networks excel at identifying complex, non-linear patterns in high-dimensional data. These models can capture relationships that would be impossible for humans to detect through traditional statistical analysis [33, 4]. However, the 'black box' nature of deep learning has historically limited its utility for scientific discovery. However, recent advances in model interpretability have begun to address this limitation, making it possible to extract human-understandable insights from trained models [20, 34, 24] – and our work at Leap Laboratories is state-of-the-art in this area.

A data-driven paradigm of this kind offers several key advantages [13, 17, 30]. By examining all available data without predetermined hypotheses, it avoids the cognitive biases that limit traditional discovery methods. Machine learning enables comprehensive pattern detection by identifying complex interactions among multiple variables that would be computationally intractable for traditional analytical approaches. Furthermore, a system that automates this analysis would provide more consistent and reproducible results, since the path from data to insight is logged and transparent, while also providing significant efficiency gains through rapid exploration of even very large, complex datasets (which currently require months of researcher time to analyse).

5 The Discovery Engine

In order for this kind of discovery to be both practical at scale and accessible to scientists without data science training, we must turn to automation. Training machine learning models well and interpreting them is a complex task that requires a great deal of expertise: practitioners in industry and academia dedicate months to manual data preparation, model training and tuning, and evaluation. Our Discovery Engine implements an end-to-end pipeline that automates this process in hours, and with the addition of our specialist interpretability methods, enables autonomous scientific discovery. Key components are as follows:

5.1 Data Ingestion and Preprocessing

The system automatically handles data cleaning, including scaling, encoding, imputation, deduplication, and outlier removal. This is particularly important because the majority of scientists are not skilled data engineers – we have found that preparing data for machine learning is far outside the expertise of most researchers, and indeed is a non-trivial task historically forming the majority of data scientists' workload. For this reason we have invested significantly in automating preprocessing as far as possible, selecting processes heuristically based on data characteristics. In practice we find this dramatically increases the time from data acquisition to discovery, with the Discovery Engine able to process many datasets with zero manual preparation.

5.2 Automated Machine Learning

A number of automated machine learning (AutoML) approaches exist, aiming to allow non-experts to achieve performance comparable to that of skilled practitioners [9, 21, 41, 16, 42]. While existing AutoML tools can reduce the need for manual model selection and hyperparameter tuning, they (i) are typically unsuitable for scientific discovery use as they rely on transfer learning from more general pre-trained models (which confuse the patterns we extract with information external to the dataset under investigation), (ii) are limited in the data structures and modalities they support, (iii) typically do not optimise models for interpretability or enable the kind of in-depth analysis necessary for knowledge discovery. The Discovery Engine therefore employs a custom AutoML system, which dynamically sizes appropriate architectures for the data; trains models

efficiently with hyperparameter search and early stopping; detects and mitigates overfit; performs a thorough performance evaluation; and ultimately selects the best-performing model. Overfit detection and mitigation is key here, and distinguishes our pattern finding from data-dredging – we test all models on holdout datasets to increase our confidence that the patterns they learn will generalise.

5.3 Automated Interpretability

Our interpretability stack employs novel techniques to extract learned patterns from the most performant model. The system distinguishes between patterns with strong empirical support present in the data and those which demonstrate extrapolation from the data by the model, categorizing findings as either 'discoveries' or 'hypotheses' respectively, and providing validatory subsets of the data for the former.

This is a key distinction: in contrast to hypothesis-generation methods, our system finds unknown patterns that *already exist* in the data (and provides subsets of the data that validate this) – putting the onus on robust data acquisition, rather than speculation reliant on an unreliable body of literature.

5.4 Report Generation

We then rank patterns based on their strength (how much they affect the variable of interest) and carefully apply LLMs to contextualize and assess the potential novelty of discovered patterns with respect to existing scientific literature. All patterns we report are statistically significant with respect to the data provided. The output of this system is then (i) a pdf report outlining and evidencing the patterns found in the dataset, ranked by strength and novelty, with reference to existing literature – also provided as a latex document, well suited for sharing in academic contexts or for forming the basis of novel publications, (ii) a dashboard allowing for interactive exploration of the patterns within the context of the dataset, (iii) code artefacts allowing for complete reproducibility of the discovery process, and (iv) the best-performing predictive model.

6 Example

6.1 The Discovery Engine in Plant Biology

Below is an example of the Discovery Engine applied to a specific problem in plant biology: that of understanding the factors that determine root architecture.

This study, conducted in collaboration with the Montpellier Institute of Plant Sciences (IPSiM), used the Discovery Engine to identify relationships between genotype, environmental variables, and early root architecture in *Arabidopsis thaliana*. *A. thaliana* is commonly used as a model organism in plant biology research due to its rapid growth cycle, genetic manipulability, and capacity to provide insights applicable to other agricultural species [23]. The investigation focused on the initial 16 days of root system

development, a critical period when root architecture begins to define how efficiently the plant will access water and nutrients [32].

The targets of interest were as follows:

- **alpha**: A weight value capturing the ratio by which the plant balances transport distance and total root growth.
- **total root length**: The sum of all root lengths per plant.
- **scaling distance to front**: The shortest distance from the observed plant architecture to the pareto front curve on a growth-transport graph.
- **mean LR angles**: The average angle of lateral roots relative to the primary root.

predicted by the following variables:

- **CO2**: Parts per million of CO2 in the room.
- **Temp**: Room temperature.
- **Genotype**: Identifies the gene mutation of the plant (or 'WT' for wild-type).
- **Nutrients**: Indicates the soil nutrients added beyond the baseline required for basic survival (where 'Mock' represents the control condition).
- **Sorbitol**: A sugar alcohol controlling osmotic stress, affecting water retention in the soil.
- **Day**: The day of the plant's development at the time of measurement (ranging between day 3 and day 15 in most cases).

Root architecture is directly correlated with critical agricultural traits such as drought resistance, nutrient uptake, and ultimately, crop yield. Understanding how roots develop under different genetic and environmental conditions is extremely important for improving agricultural practices – especially as we face increasingly unstable climates [6, 22, 37, 32].

After processing, the dataset used in this study contained approximately 700 samples, over 20 genotypes and 50 different nutrient treatments across different environmental settings. The challenge scientists would typically face is finding combinations of these inputs that work together to optimise root structures. This is a non-trivial task, since existing analysis methods struggle to identify combinatorial, non-linear patterns. Our collaborators at IPSiM told us that they would typically spend 'months scrolling in excel' to make sense of this data – in contrast to our automated system, which identified these relationships in less than an hour.

The Discovery Engine extracted 39 patterns describing relationships between the experimental variables and the targets of interest. Of these, 18 appear to be unexplored in existing literature. A publication describing the most interesting of these is currently underway, so to avoid duplication we include only a single pattern here.

In Figure 1, we present a visualisation of this pattern as a violin plot. Each violin represents the distribution of a single variable, shown on the y axis, under different

sets of conditions (or rules) extracted by our system, shown on the x axis. Above each violin, we show the p-value (p), the mean of the target value (μ) in the data subset (also denoted by a horizontal dotted line), and the number of samples (n) in that subset. Below the plot, we include a brief LLM-generated statement contextualising the pattern with respect to existing knowledge.

The full output of the Discovery Engine includes this information for every pattern found, along with a dashboard allowing for interactive exploration of the patterns within the context of the dataset.

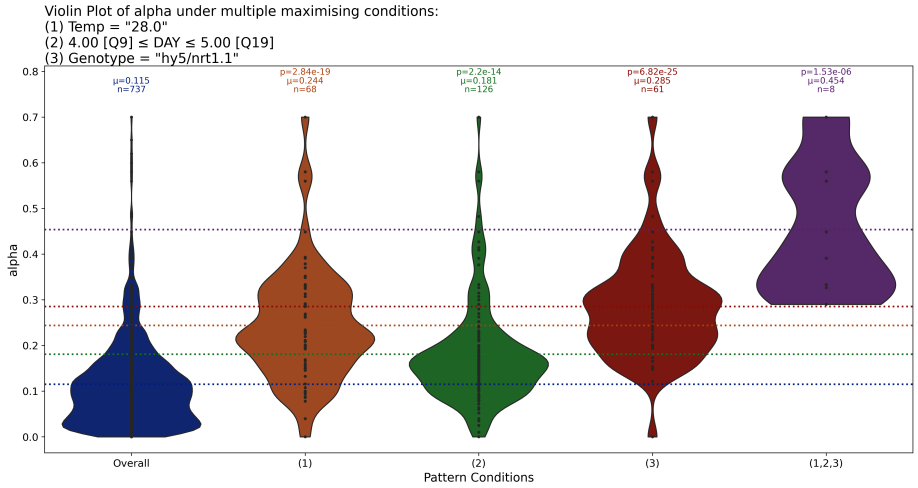

Figure 1: The *alpha* value tends to be higher for the hy5/nrt1.1 genotype at warm temperatures and early stages of development.

> 'This specific pattern relating the hy5/nrt1.1 genotype, temperature, and development stage to the alpha weight is not established in existing literature. Related work has shown that the HY5 and NRT1.1 genes are involved in regulating root growth and nitrogen uptake respectively, which could help explain the observed relationship with the alpha value that balances transport distance and total root growth. However, the specific interaction of this double mutant with temperature and development stage is a novel finding that warrants further investigation.' – LLM

On the far left of Figure 1, in the first violin, we see the full distribution of *alpha* values over the entire dataset. Condition (1) 'Temp = 28.0' significantly increases the mean *alpha*, shown in the second violin. The third violin shows the distribution of *alpha* under condition (2), where the measurement is taken on day 4 or 5. The distribution of alpha in samples where condition (3), 'Genotype = hy5/nrt1.1', is shown in the fourth violin – this also significantly increases the mean *alpha* value compared to the overall distribution. The fifth violin shows that **when these conditions are present together, the mean *alpha* value increases far more than when any condition is present alone**. This demonstrates a combinatorial effect, that would be extremely hard to identify via

manual analysis unless the practitioner were actively looking for it. We have found a novel insight into how environmental variables, plant age and genotype affect root structure, that would otherwise have been missed.

Plant scientists at IPSiM are now able to test many more variables in a single experiment than they could previously, massively increasing iteration speed and efficiency – because with the Discovery Engine, they are able to easily and automatically extract meaningful patterns from the results – rather than spending months on laborious manual exploration of the data, or limiting their enquiry to a few pre-conceived hypotheses.

Other collaborations of this nature have yielded more valuable insights. Over the coming months, working closely with our scientific partners, we plan to publish this study and a number of other applications of our Discovery Engine – in domains spanning meteorology, neuroscience, genomics, materials and more.

7 Conclusion

The data-driven discovery paradigm represents a fundamental shift in scientific methodology. By removing hypothesis-driven constraints, we accelerate discovery rates and reduce the impact of cognitive biases – and with automation, we make this capability accessible to all scientists. However, successful and widespread implementation requires changes in research practices, including increased emphasis on comprehensive data collection and sharing. The key limitation of data-driven discovery, is of course, data. The Discovery Engine is entirely reliant on data quality (and to a lesser extent, quantity) – bad (noisy, missing, biased, error-prone) data makes it extremely difficult for models (and humans!) to learn anything useful. For this reason, we see the role of scientists in a post-discovery-engine world as one primarily of skilled data collectors – and naturally, as directors of enquiry. The Discovery Engine is not opinionated about *which* phenomena to investigate. It asks only for data of sufficient quality, and provides a superhuman lens through which we can make sense of it.

References

[1] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604): 452–454, 2016. doi: 10.1038/533452a. URL https://www.nature.com/articles/533452a.

[2] A Bate, M Lindquist, and IR Edwards. The application of knowledge discovery in databases to post-marketing drug safety: example of the who database. *Fundamental & clinical pharmacology*, 22(2):127–140, 2008.

[3] Nicholas Bloom, Charles I Jones, John Van Reenen, and Michael Webb. Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144, 2020. doi: 10.1257/aer.20180338. URL https://www.aeaweb.org/articles?id=10.1257/aer.20180338.

[4] Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559 (7715):547–555, 2018. doi: 10.1038/s41586-018-0337-2. URL https://www.nature.com/articles/s41586-018-0337-2.

[5] Tingting Chen, Srinivas Anumasa, Beibei Lin, Vedant Shah, Anirudh Goyal, and Dianbo Liu. Auto-bench: An automated benchmark for scientific discovery in llms. *arXiv preprint*, 02 2025. doi: 10.48550/arXiv.2502.15224.

[6] Yinglong Chen, Zed Rengel, Jairo Palta, and Kadambot HM Siddique. Efficient root systems for enhancing tolerance of crops to water and phosphorus limitation. *Indian Journal of Plant Physiology*, 23:689–696, 2018.

[7] Johan SG Chu and James A Evans. Slowed canonical progress in large fields of science. *Proceedings of the National Academy of Sciences*, 118(41):e2021636118, 2021. doi: 10.1073/pnas.2021636118. URL https://www.pnas.org/doi/10.1073/pnas.2021636118.

[8] Daniele Fanelli. Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3):891–904, 2012. doi: 10.1007/s11192-011-0494-7. URL https://link.springer.com/article/10.1007/s11192-011-0494-7.

[9] Luis Ferreira, Andre Pilastri, Carlos Manuel Martins, Pedro Miguel Pires, and Paulo Cortez. A Comparison of AutoML Tools for Machine Learning, Deep Learning and XGBoost. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, jul 18 2021.

[10] Annie Franco, Neil Malhotra, and Gabor Simonovits. Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6251):1502–1505, 2014. doi: 10.1126/science.1255484. URL https://www.science.org/doi/10.1126/science.1255484.

[11] William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus. Knowledge discovery in databases: An overview. *AI Magazine*, 13 (3):57–70, 1992. doi: https://doi.org/10.1609/aimag.v13i3.1011. URL https://onlinelibrary.wiley.com/doi/abs/10.1609/aimag.v13i3.1011.

[12] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an ai co-scientist, 2025. URL https://arxiv.org/abs/2502.18864.

[13] Tony Hey, Stewart Tansley, and Kristin Tolle. *The fourth paradigm: Data-intensive scientific discovery*. Microsoft Research, 2009. URL https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

[14] John PA Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005. doi: 10.1371/journal.pmed.0020124. URL https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124.

[15] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[16] Shubhra Kanti Karmaker ("Santu"), Md. Mahadi Hassan, Micah J. Smith, Lei Xu, Chengxiang Zhai, and Kalyan Veeramachaneni. Automl to Date and Beyond: Challenges and Opportunities. *ACM Computing Surveys*, 54(8):1–36, oct 4 2021.

[17] Rob Kitchin. Big data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):2053951714528481, 2014. doi: 10.1177/2053951714528481. URL https://journals.sagepub.com/doi/10.1177/2053951714528481.

[18] Adithya Kulkarni, Fatimah Alotaibi, Xinyue Zeng, Longfeng Wu, Tong Zeng, Barry Yao, Minqian Liu, Shuaicheng Zhang, Lifu Huang, and Dawei Zhou. Scientific hypothesis generation and validation: Methods, datasets, and future directions. *arXiv preprint*, 05 2025. doi: 10.48550/arXiv.2505.04651.

[19] Zhehui Liao, Maria Antoniak, Inyoung Cheong, Evie Yu-Yen Cheng, Ai-Heng Lee, Kyle Lo, Joseph Chee Chang, and Amy X. Zhang. Llms as research tools: A large scale survey of researchers' usage and perceptions, 2024. URL https://arxiv.org/abs/2411.05025.

[20] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017. URL https://dl.acm.org/doi/10.5555/3295222.3295230.

[21] Lindauer M. Hands-On Automated Machine Learning Tools: Auto-Sklearn and Auto-PyTorch. *AMIR@ECIR*, 2019.

[22] Riya Mazumdar, Surajeet Konwar, Tridisha Borgohain, Dashami Das, Sewashri Das, Ankur Jyoti Dutta, Smita Doley, Pallabi Dutta, and Florina Rabha. Drought stress in plants: A review on morphological, physiological and biochemical alterations with special reference to drought stress tolerance strategies. *Ecology, Environment & Conservation (0971765X)*, 30(4), 2024.

[23] Elliot M Meyerowitz. Arabidopsis thaliana. *Annual review of genetics*, 21(1): 93–111, 1987.

[24] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Self-published, 3rd edition, 2025. ISBN 978-3-911578-03-5. URL https://christophm.github.io/interpretable-ml-book/.

[25] Miryam Naddaf. How are researchers using ai? survey reveals pros and cons for science. *Nature*, 2025. URL https://api.semanticscholar.org/CorpusID:276113654.

[26] Brian A Nosek, Jeffrey R Spies, and Matt Motyl. Scientific utopia: Ii. restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6):615–631, 2012. doi: 10.1177/1745691612459058. URL https://journals.sagepub.com/doi/10.1177/1745691612459058.

[27] Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015. doi: 10.1126/science.aac4716. URL https://www.science.org/doi/10.1126/science.aac4716.

[28] Margit E Oswald and Stefan Grosjean. Confirmation bias. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, 79:83, 2004.

[29] Michael Park, Erin Leahey, and Russell J Funk. Papers and patents are becoming less disruptive over time. *Nature*, 613(7942):138–144, 2023. doi: 10.1038/s41586-022-05543-x. URL https://www.nature.com/articles/s41586-022-05543-x.

[30] Wolfgang Pietsch. *On the Epistemology of Data Science: Conceptual Tools for a New Inductivism*, volume 148 of *Philosophical Studies Series*. Springer, 2021. ISBN 978-3030864415. URL https://link.springer.com/book/10.1007/978-3-030-86442-2.

[31] Manuela Fernández Pinto. Methodological and cognitive biases in science: Issues for current research and ways to counteract them. *Perspectives on Science*, 31 (5):535, 2023. URL https://direct.mit.edu/posc/article/31/5/535/115648/.

[32] Alok Ranjan, Ragini Sinha, Sneh L Singla-Pareek, Ashwani Pareek, and Anil Kumar Singh. Shaping the root system architecture in plants for adaptation to drought stress. *Physiologia plantarum*, 174(2):e13651, 2022.

[33] Markus Reichstein, Gustavo Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019. doi: 10.1038/s41586-019-0912-1. URL https://www.nature.com/articles/s41586-019-0912-1.

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016. doi: 10.1145/2939672.2939778. URL https://dl.acm.org/doi/10.1145/2939672.2939778.

[35] Peter J. Richardson, Bruce W. S. Robinson, Daniel P. Smith, and Justin Stebbing. The ai-assisted identification and clinical efficacy of baricitinib in the treatment of covid-19. *Vaccines*, 10(6):951, 2022. doi: 10.3390/vaccines10060951. URL https://doi.org/10.3390/vaccines10060951.

[36] Marta Serra-Garcia and Uri Gneezy. Nonreplicable publications are cited more than replicable ones. *Science Advances*, 7(21):eabd1705, 2021. doi: 10.1126/sciadv.abd1705.

[37] Devesh Shukla, Prabodh Kumar Trivedi, and Shivendra Sahi. A simple protocol for mapping the plant root system architecture traits. *Journal of Visualized Experiments (JoVE)*, 192(192):e64876, 2023.

[38] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011. doi: 10.1177/0956797611417632. URL https://journals.sagepub.com/doi/10.1177/0956797611417632.

[39] Justin Sybrandt, Micheal Shtutman, and Ilya Safro. Large-scale validation of hypothesis generation systems via candidate ranking. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1494–1503, 2018. doi: 10.1109/BigData.2018.8622637.

[40] Rick P Thomas, Michael R Dougherty, Amber M Sprenger, and J Isaiah Harbison. Implications of cognitive load for hypothesis generation and probability judgment. *Frontiers in Psychology*, 2:129, 2011. doi: 10.3389/fpsyg.2011.00129.

[41] Anh Truong, Austin Walters, Jeremy Goodsitt, Keegan Hines, C. Bayan Bruss, and Reza Farivar. Towards Automated Machine Learning: Evaluation and Comparison of AutoML Approaches and Tools. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1471–1479. IEEE, 11 2019.

[42] Doris Xin, Eva Yiwei Wu, Doris Jung-Lin Lee, Niloufar Salehi, and Aditya Parameswaran. Whither AutoML? Understanding the Role of Automation in Machine Learning Workflows. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16. ACM, may 6 2021.

[43] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32, 2024.

[44] Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. A comprehensive survey of scientific large language models and their applications in scientific discovery. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8783–8817, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.498. URL https://aclanthology.org/2024.emnlp-main.498/.