Review

Generative Metascience: A Review of AI as the Next Scientific Instrument and the Emerging Paradigm of Algorithmic Discovery

Shawn Ray *

Posted Date: 14 July 2025

doi: 10.20944/preprints202507.0417.v4

Keywords: generative metascience; algorithmic discovery; AI-driven scientific instrumentation; hypothesis generation and testing; research automation and workflow optimization

 doi:10.20944/preprints202507.0417.v4

Review

Generative Metascience: A Review of AI as the Next Scientific Instrument and the Emerging Paradigm of Algorithmic Discovery

Shawn Ray

Carnegie Mellon University, Pittsburgh, United States; shawnray5699@gmail.com

Abstract

This review introduces Generative Metascience, a comprehensive framework for understanding how artificial intelligence (AI) transforms scientific discovery. We synthesize historical milestones and case studies across genomics, astronomy, materials science, and social sciences to illustrate AI's evolution from a research instrument into an autonomous co-investigator. Our analysis is structured around two core themes: AI-enabled data collection and analysis, and AI-driven hypothesis generation and testing. This dual focus highlights the iterative interplay between data-driven analytics and hypothesis-driven inquiry, showing how AI tools can simultaneously generate and evaluate scientific hypotheses. Key insights reveal AI-accelerated breakthroughs, such as automated protein-folding and materials design, and the rise of self-driving laboratories, which signal a shift from traditional inquiry toward an algorithmic discovery paradigm. By synthesizing data-driven pattern recognition with AI-facilitated hypothesis generation across disciplines, this review addresses a critical research gap, showing that AI has begun to automate hypothesis formation and serves as a meta-technology that is redefining scientific epistemology. We highlight urgent implications for future research, including the development of hybrid AI-human workflows and robust metrics for machine-generated insights. For policy, we articulate the need for greater transparency, open data standards, and interdisciplinary funding initiatives. For practice, we advocate for retraining researchers and updating curricula for AI-integrated labs. By articulating these developments and their contributions, this paper charts a roadmap for responsibly harnessing AI's potential and guiding the scientific community as it navigates AI's evolving role in a new era of discovery.

Keywords: generative metascience; algorithmic discovery; AI-driven scientific instrumentation; hypothesis generation and testing; research automation and workflow optimization

1. Introduction: Establishing the Context and Rationale

1.1. The Broader Landscape: Background and Significance

In 2020, DeepMind's AlphaFold achieved a groundbreaking milestone by solving the protein folding problem with remarkable accuracy, a challenge that had persisted for over five decades (Jumper et al., 2021). This success, powered by advanced machine learning, exemplifies AI's transformative potential in scientific discovery. Beyond protein folding, AI has made significant contributions across various scientific domains. In astronomy, machine learning models have been instrumental in detecting exoplanets, with projects like ExoMiner using deep learning to validate 301 exoplanets from Kepler mission data (Valizadegan et al., 2022). In drug discovery, AI has accelerated the identification of new therapeutic compounds; for instance, Exscientia has utilized AI to design molecules that have entered clinical trials, significantly reducing the traditional timeline for drug development (Philippidis, 2023). These examples underscore AI's versatility and its capacity to address complex scientific challenges.

 doi:10.20944/preprints202507.0417.v4

The volume of scientific literature is growing exponentially, with recent studies indicating a doubling time of approximately 17.3 years (Bornmann et al., 2021). This rapid expansion underscores the critical role of metascience in enhancing the efficiency and reliability of research practices, ensuring that scientific inquiry remains robust and impactful amidst mounting complexity.

Artificial intelligence, with its ability to process vast datasets and discern intricate patterns, is becoming a pivotal force in metascience. AI can automate mundane tasks like data cleaning, generate novel hypotheses via pattern recognition and predictive modeling, optimize experimental designs, and interpret results by highlighting significant trends and anomalies (Jordan & Mitchell, 2015a). These functionalities not only expedite scientific discovery but also challenge traditional paradigms by offering innovative methods for hypothesis generation and validation. A notable example is in drug discovery, where AI has facilitated the creation of new therapeutic compounds; for instance, Insilico Medicine has used AI to design drugs that have progressed to clinical trials, markedly reducing development timelines and costs (Zhavoronkov et al., 2019).

This review introduces Generative Metascience, a framework that emphasizes AI's, particularly generative models', capacity to not only analyze existing data but also to autonomously generate new hypotheses and propel scientific discovery. This approach underscores AI's dual function as both an analytical instrument and an independent co-investigator in research. By synthesizing AI's applications across various scientific fields, from genomics to astronomy, we seek to elucidate how AI is fundamentally altering the terrain of scientific inquiry.

1.2. The Core Problem: Identifying the Critical Gap or Controversy

Despite AI's widespread adoption in scientific research, there is a pressing need for a comprehensive review that assesses its overarching role as a scientific instrument. Existing literature tends to concentrate on specific applications, such as AI in drug discovery or AI in astronomy, offering detailed insights into particular domains but neglecting the broader, interdisciplinary implications of AI (Ball & Brunner, 2010; Zhavoronkov et al., 2019). A holistic analysis that examines AI's influence on the entire scientific process, encompassing methodological shifts, ethical dilemmas, and the practical challenges of incorporating AI into conventional research practices, is conspicuously absent. This deficiency impedes our capacity to fully harness AI's potential while addressing critical risks, including data biases and the interpretability of AI models (Jobin et al., 2019).

1.3. Delineating the Review's Focus and Approach

Generative Metascience refers to an AI-driven framework in which artificial intelligence not only analyzes existing scientific data but also autonomously formulates, prioritizes, and tests novel hypotheses, effectively acting as both a research instrument and co-investigator in the scientific discovery process. This review aims to fill this critical gap by offering a comprehensive analysis of AI's multifaceted applications in scientific research, evaluating its efficacy, and delving into the nascent paradigm of algorithmic discovery. Uniquely, this review integrates AI's functions in both data collection and analysis and in hypothesis generation and testing, presenting a unified framework that encompasses AI's complete influence on the scientific endeavor. By focusing on this duality, we elucidate the dynamic between data-centric analytics and hypothesis-oriented research, illustrating how AI can concurrently formulate and appraise scientific hypotheses. Drawing from a synthesis of findings across various disciplines, we identify exemplary practices for incorporating AI into research methodologies and pinpoint areas necessitating further scholarly attention. Additionally, this review will address the policy implications of AI's integration into science, including recommendations for funding initiatives and educational reforms to support the development of AI-driven research ecosystems. This balanced synthesis is intended to inform and guide researchers, policymakers, and funding bodies in navigating the complexities of AI integration in science.

 doi:10.20944/preprints202507.0417.v4

1.4. A Roadmap for the Reader: Structure of the Article

This review is structured as follows: Section 2, "Thematic Synthesis and Critical Analysis of the Literature," encompasses five subsections that explore AI's role in scientific discovery. Subsection 2.1, "Historical and Conceptual Foundations," provides a historical overview of AI's development in science and clarifies key terminology. Subsection 2.2, "The Methodological Canvas: Approaches to Research in the Field," examines the diverse methodologies used to study AI in scientific contexts, including their strengths, weaknesses, and emerging innovations. Subsections 2.3 and 2.4 delve into the two principal themes: "Thematic Deep Dive 1" focuses on AI's contributions to data collection and analysis, while "Thematic Deep Dive 2" addresses AI's role in hypothesis generation and testing. Subsection 2.5, "Cross-Thematic Analysis: Interconnections and Contrasting Perspectives," synthesizes the interactions between these themes and their implications for the paradigm of algorithmic discovery, Section 3, "Discussion, Implications, and Future Trajectories," outlines the theoretical and practical implications of AI-driven science, proposes ethical strategies for AI integration, and highlights key areas for future research. Finally, Appendix A is appended to this paper to demonstrate the methodologies used to compile and analyze the information presented in this paper.

2. Thematic Synthesis and Critical Analysis of the Literature

2.1. Historical and Conceptual Foundations

This section establishes the historical and theoretical groundwork for understanding the emergence of Generative Metascience, a framework that positions AI as both an analytical instrument and an autonomous agent capable of generating novel scientific hypotheses and driving independent research. By tracing AI's evolution in scientific discovery, we illustrate how each advancement has contributed to this paradigm, highlighting the transition from data-driven analysis to generative and autonomous scientific inquiry.

2.1.1. The Genesis of the Field: Seminal Works and Key Milestones

The integration of AI into scientific discovery began in the 1960s with DENDRAL, developed at Stanford University (Buchanan & Feigenbaum, 1981). As the first expert system, DENDRAL assisted organic chemists in identifying unknown molecules by analyzing mass spectrometry data, marking an early step toward automating hypothesis formation. This pioneering work demonstrated AI's potential to augment human reasoning, laying the foundation for its role in Generative Metascience by enabling structured, rule-based hypothesis generation.

In the 1990s and 2000s, the advent of machine learning techniques, such as support vector machines and random forests, expanded AI's capabilities. These methods were particularly impactful in bioinformatics, facilitating tasks like gene expression analysis and protein classification (Baldi & Brunak, 2001). By extracting meaningful patterns from large datasets, machine learning enabled data-driven hypothesis generation, a critical precursor to the generative aspects of modern AI systems.

The deep learning revolution of the 2010s marked a significant leap, with neural networks achieving breakthroughs across diverse scientific domains. In 2017, artificial neural networks addressed the quantum many-body problem, a longstanding challenge in physics (Iten et al., 2020). In 2020, DeepMind's AlphaFold solved the 50-year-old protein folding problem with unprecedented accuracy, predicting protein structures from amino acid sequences (Jumper et al., 2021). This achievement not only accelerated biological research but also exemplified AI's ability to integrate data analysis with hypothesis generation, a core tenet of Generative Metascience.

The rise of generative AI models further advanced this paradigm. Generative adversarial networks (GANs) and variational autoencoders (VAEs) have been used to design novel molecules and materials, enabling AI to propose new scientific entities that drive hypothesis formation (Mi et

 doi:10.20944/preprints202507.0417.v4

al., 2018). For instance, in materials science, AI has identified promising battery material candidates, significantly reducing the time required for traditional trial-and-error methods (Lv et al., 2022).

In recent years, large language models (LLMs) and foundation models have emerged as transformative tools in scientific discovery. Models like GPT-4 have demonstrated capabilities in generating scientific text, code, and hypotheses, aligning closely with the generative metascience framework. In 2024, the AI Scientist framework enabled LLMs to conduct research autonomously, from idea generation to paper writing (Lu et al., 2024). A notable example is Sakana AI's AI Scientist-v2, which in 2025 autonomously generated a hypothesis, designed experiments, and produced a paper accepted at a top machine learning conference, though it was later withdrawn due to ethical concerns (Yamada et al., 2025). These milestones highlight AI's growing autonomy, positioning it as a proactive collaborator in scientific inquiry.

These developments reflect key trends: a progression from narrow, task-specific AI to general, autonomous systems; an expansion in the scale and complexity of problems addressed; and a shift from supportive to proactive roles in science. Each milestone has advanced AI's generative capabilities, aligning with the principles of Generative Metascience by enabling AI to propose and test novel scientific ideas across disciplines.

2.1.2. The Evolution of Core Concepts and Theories

The conceptual evolution of AI in science mirrors its historical milestones, progressing from rule-based systems to data-driven, deep learning, and generative models, each addressing limitations of prior approaches and contributing to the framework of Generative Metascience. Early systems like DENDRAL relied on manually encoded rules, limiting their generalizability and requiring extensive expert input. This constrained their ability to generate novel hypotheses beyond predefined knowledge, a significant limitation for scientific discovery.

The shift to machine learning in the 1990s introduced data-driven methods, such as decision trees and ensemble methods, which learned from examples and adapted to new data (Rifkin, 2002). These approaches enabled AI to handle complex datasets, facilitating pattern recognition that informed hypothesis generation in fields like bioinformatics. This marked an early step toward generative capabilities, as AI began to identify patterns that could inspire new scientific inquiries.

The deep learning revolution of the 2010s overcame the limitations of traditional machine learning by leveraging multi-layered neural networks to model complex relationships (Sejnowski, 2018). This enabled breakthroughs in image analysis for astronomy, sequence prediction in genomics, and natural language processing for scientific literature. However, deep learning's "black box" nature posed challenges for scientific validation, prompting the development of interpretability methods like SHAP and LIME to enhance trust in AI-driven findings (Xu et al., 2019).

Generative AI models, such as GANs and VAEs, further expanded AI's role by enabling the creation of novel scientific entities, such as molecules and experimental designs (Mi et al., 2018). These models directly support Generative Metascience by generating hypotheses that push scientific boundaries, moving beyond analysis to innovation.

The emergence of LLMs and foundation models represents the latest advancement, enabling AI to generate scientific text, code, and hypotheses. For instance, LLMs have been used to draft research proposals and analyze literature, streamlining the scientific process (Liang et al., 2024). The AI Scientist framework exemplifies this, integrating data analysis, hypothesis generation, and experiment design into an autonomous research cycle (Lu et al., 2024). This shift challenges traditional scientific workflows, raising questions about the role of human researchers and the nature of scientific creativity.

Each stage of this evolution has addressed prior limitations: from the rigidity of rule-based systems to the scalability of machine learning, the complexity handling of deep learning, and the generative and autonomous capabilities of modern AI. This progression underscores AI's transformative potential in redefining the scientific method, with Generative Metascience providing a framework to understand and guide this transition. Figure 1 illustrates this conceptual evolution as

 doi:10.20944/preprints202507.0417.v4

a chronological timeline of key AI milestones in scientific discovery, from the rule-based reasoning of DENDRAL in the 1960s to the fully autonomous AI Scientist of 2025. Together, these developments underscore how each paradigm shift—rule-based, data-driven, deep learning, and generative models—builds toward the framework of Generative Metascience (see Figure 1).

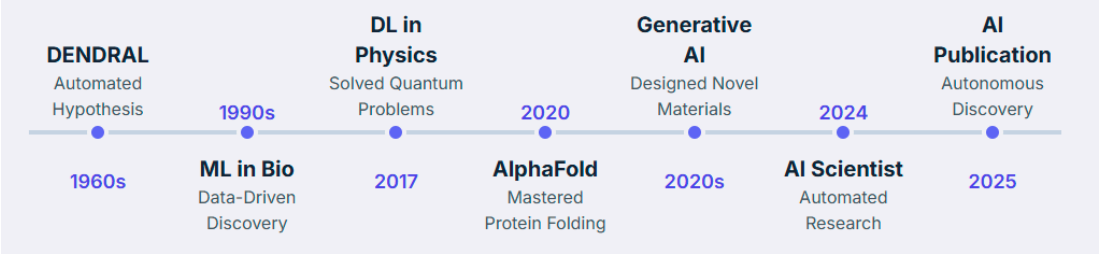

Figure 1. AI in Science Over Time.

2.2. The Methodological Canvas: Approaches to Research in the Field

The exploration of artificial intelligence (AI) as a transformative scientific instrument within the framework of Generative Metascience relies on a diverse set of research methodologies. These approaches, spanning qualitative and quantitative paradigms, investigate AI's integration into scientific workflows, its effectiveness, and its broader implications for the scientific process. This section critically evaluates these methodologies, providing specific examples from the literature, addressing their strengths and limitations, and highlighting emerging innovations that enhance the study of AI in science. Figure 2 presents a methodological canvas that maps how simulations, case studies, and experimental studies each contribute to broader scientific implications, effectiveness & validation, and AI integration within scientific workflows. See Figure 2 for a visual summary of these interrelated approaches.

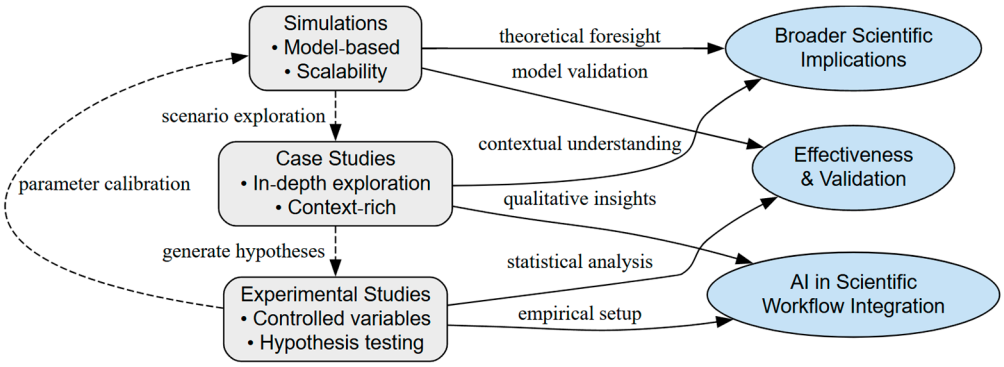

Figure 2. Interaction of Core AI Methodologies.

2.2.1. Predominant Research Methodologies: A Critical Overview

The methodologies employed to study AI's role in scientific discovery are varied, each offering distinct perspectives on how AI reshapes research practices. Below, we outline the primary methodologies, supported by specific examples and peer-reviewed references, to illustrate their application in the context of Generative Metascience.

- **Case Studies**: Case studies provide in-depth analyses of specific AI applications, offering rich, contextual insights into their integration with scientific workflows. A prominent example is DeepMind's AlphaFold, which solved the protein folding problem with unprecedented accuracy, demonstrating AI's capacity to address complex scientific challenges (Jumper et al., 2021). Another significant case is the development of DSP-1181, the first AI-designed drug to enter clinical trials, created through a collaboration between Exscientia and Sumitomo Dainippon Pharma for obsessive-compulsive disorder treatment. This project completed its

 doi:10.20944/preprints202507.0417.v4

exploratory research phase in under 12 months, compared to the traditional 4-6 years, showcasing AI's potential to accelerate drug discovery (Burki, 2020). These cases highlight AI's role in generating and testing hypotheses, aligning with the principles of Generative Metascience.

- **Experimental Studies**: Experimental studies compare AI-driven methods with traditional approaches, providing rigorous evidence of AI's efficacy. For instance, Granda et al. (2018) demonstrated that AI-optimized chemical reaction conditions outperformed human-designed methods, achieving higher efficiency in organic synthesis (Granda et al., 2018). Such studies validate AI's practical utility in scientific tasks, supporting its role as a generative tool in research.
- **Surveys and Interviews**: These methods capture scientists' perceptions of AI, revealing adoption barriers such as lack of training or concerns about model transparency. A survey by the Center for Science, Technology and Environmental Policy Studies at Arizona State University found that while scientists recognize AI's potential to enhance research, many express concerns about its impact on scientific integrity and the need for ethical guidelines (Z. Chen et al., 2024). Similarly, a Pew Research Center survey reported that 52% of Americans are more concerned than excited about AI's role in daily life, reflecting broader societal apprehensions that influence scientific adoption (Zhang & Dafoe, 2019).
- **Data Mining and Bibliometric Analyses**: These approaches identify trends in AI's application across disciplines by analyzing large datasets of publications. Rahman et al. (2024) conducted a bibliometric analysis of AI in medical diagnoses, noting a significant increase in machine learning and deep learning usage as such, underscoring AI's growing influence (Rahman et al., 2024). These analyses provide a macro-level perspective on AI's integration into scientific research.
- **Simulations**: Simulations test AI in virtual environments, offering cost-effective ways to evaluate its predictive capabilities. Raccuglia et al. (2016) used machine learning to simulate and accelerate the discovery of new materials, demonstrating AI's ability to explore complex systems where physical experiments are impractical (Raccuglia et al., 2016). Such simulations support hypothesis generation, a key aspect of Generative Metascience.
- **Theoretical Modeling**: Theoretical models develop frameworks to understand AI's role in science. Jordan and Mitchell (2015) proposed a model for how AI can reshape hypothesis generation and testing, providing a conceptual foundation for studying AI's impact on scientific discovery (Jordan & Mitchell, 2015). These models guide empirical research but require validation to ensure practical relevance.

The prevalence of case studies and experimental studies reflects their ability to provide detailed insights into AI's transformative applications and empirical evidence of its advantages. Surveys and bibliometric analyses offer broader perspectives on research trends and community perceptions, while simulations and theoretical modeling enable exploration in controlled or conceptual settings. This methodological diversity ensures a comprehensive understanding of AI's role in advancing scientific inquiry.

2.2.2. Strengths, Weaknesses, and the Rise of Innovative Methods

Each methodology offers distinct strengths and faces inherent limitations, shaping their suitability for studying AI in science. Below, we critically analyze these aspects, supported by examples, and discuss emerging innovations that address these limitations. Table 1 provides a concise overview of these methodologies, highlighting each approach's key strengths, notable weaknesses, and a representative example. See Table 1 for details.

 doi:10.20944/preprints202507.0417.v4

Table 1. AI Methodologies.

	Strengths	Weaknesses	Example
Case Studies	Provide rich, contextual insights into AI applications, capturing complex interactions between technology and science.	Often context-specific, limiting generalizability; prone to selection bias, overrepresenting successful cases.	AlphaFold's protein folding solution
Experimental Studies	Offer rigorous evidence of causality and quantitative comparisons, ideal for assessing AI's effectiveness.	Controlled settings may not reflect real-world complexities; scaling results to practical applications can be challenging.	AI-optimized chemical synthesis
Surveys and Interviews	Excel at capturing subjective experiences and human perspectives, essential for understanding AI adoption barriers.	Susceptible to biases like social desirability or sampling issues; qualitative data interpretation can be subjective.	Scientists' concerns about AI ethics
Data Mining and Bibliometric Analyses	Handle large datasets efficiently, providing objective, macro-level insights into trends and patterns.	Depend on data quality; risk of misinterpretation if analytical methods are not robust.	AI publication trends in drug discovery
Simulations	Offer flexibility and cost-effectiveness for testing AI in impractical scenarios.	Validity hinges on accurate underlying models; unrealistic assumptions can undermine results.	Material discovery simulations
Theoretical Modeling	Provide structured frameworks for understanding AI's role, guiding empirical research.	Risk being speculative without empirical validation, disconnecting from practical applications.	AI's impact on hypothesis generation

Critical Analysis of Methodological Blind Spots

While these methodologies collectively advance our understanding of AI in science, they exhibit notable blind spots. Case studies, such as those on AlphaFold, often focus on high-profile successes, potentially overlooking less successful applications that could reveal critical limitations of AI systems. Experimental studies, like Granda et al. (2018), may prioritize controlled environments, missing the nuanced challenges of real-world scientific contexts, such as interdisciplinary

 doi:10.20944/preprints202507.0417.v4

complexities or data variability. Surveys and interviews, as seen in the ASU SciOPS study, may suffer from response biases, particularly if participants are hesitant to express critical views about AI due to its perceived transformative potential. Data mining and bibliometric analyses, while comprehensive, may miss emerging trends not yet reflected in publication databases, limiting their ability to capture cutting-edge developments. Simulations, such as those by Raccuglia et al. (2016), rely on model assumptions that may not fully represent real-world phenomena, potentially leading to overoptimistic predictions. Theoretical models, like those proposed by Jordan and Mitchell (2015), risk being overly abstract without sufficient empirical grounding, which can hinder their practical utility.

Emerging Innovative Methods

To address these limitations, innovative methodologies are emerging to enhance the rigor, transparency, and ethical integration of AI in scientific research. AI-powered literature reviews, such as those using natural language processing (NLP) to analyze thousands of papers, enable rapid synthesis of research trends, as demonstrated by tools like Semantic Scholar (Kinney et al., 2025). Automated meta-analyses leverage AI to conduct systematic reviews with reduced human bias, improving efficiency in synthesizing evidence across studies (Harrer et al., 2019). Reproducibility frameworks use AI to verify data integrity and methodological consistency, enhancing trust in AI-driven findings (Haibe-Kains et al., 2020). Social media sentiment analysis, such as analyses of X posts, gauges the scientific community's reactions to AI advancements, providing real-time insights into adoption trends (Qi et al., 2024). Ethical assessment frameworks, like those proposed by Jobin et al. (2019), offer guidelines for addressing bias, privacy, and accountability in AI applications. These innovations align with Generative Metascience by fostering a more robust and responsible approach to studying AI's role in science.

2.3. Thematic Deep Dive 1

This section examines artificial intelligence (AI) as a cornerstone of scientific data collection and analysis, a key pillar of Generative Metascience. By leveraging machine learning and deep learning, AI enables researchers to process vast, complex datasets with unprecedented efficiency and accuracy, driving discoveries across disciplines such as astronomy, genomics, and particle physics. Structured into synthesis, consensus/controversy, and critique, this section highlights AI's transformative impact, addresses ongoing debates, and identifies methodological gaps requiring further exploration.

2.3.1. Synthesis of Key Findings and Supporting Evidence

AI has reshaped scientific data collection and analysis by automating complex tasks and uncovering patterns that were previously unattainable. In astronomy, AI manages massive datasets generated by modern telescopes. The Vera C. Rubin Observatory, for instance, is expected to produce 0.5 exabytes of data over its 10-year survey, equivalent to 50,000 times the Library of Congress's book collection (Thomas et al., 2020). Neural networks have achieved high accuracy in classifying galaxy morphologies; a seminal study by Banerji et al. (2010) used neural networks on Sloan Digital Sky Survey data, achieving 98% accuracy comparable to human experts (Banerji et al., 2012). Similarly, AI-driven exoplanet detection has reached 96% accuracy using Kepler mission data, as demonstrated by Valizadegan et al. (2022) in the ExoMiner project (Valizadegan et al., 2022).

In genomics, AI accelerates the analysis of intricate genetic data, advancing personalized medicine and disease research. DeepVariant, a deep learning-based variant caller developed by Google, significantly improves the accuracy of identifying genetic variants, outperforming traditional tools without requiring specialized domain knowledge (Poplin, Chang, et al., 2018; Poplin, Varadarajan, et al., 2018). SpliceAI, a 32-layer deep neural network, predicts splicing from DNA sequences with up to 95% accuracy, aiding in identifying cryptic splicing variants linked to

 doi:10.20944/preprints202507.0417.v4

neurodevelopmental disorders (Jaganathan et al., 2019). These tools exemplify AI's ability to handle the scale and complexity of modern sequencing data.

In particle physics, AI is critical for analyzing petabytes of data from high-energy collisions at the Large Hadron Collider (LHC). The CMS collaboration employs AI to detect anomalous jets, enhancing sensitivity to new physics phenomena (C. M. S. Collaboration, 2021). Similarly, ATLAS uses deep learning for precise identification of b-hadrons and other particles (A. Collaboration, 2020). These applications underscore AI's role in managing data complexity and volume.

AI's impact extends to chemistry, where it predicts molecular properties, and materials science, where it accelerates novel material discovery. In environmental science, AI models climate data and predicts natural disasters, enabling proactive responses. By automating routine tasks and revealing hidden patterns, AI frees researchers to focus on innovative hypotheses, aligning with Generative Metascience's emphasis on advancing scientific inquiry. Table 2 summarizes key AI applications across major scientific fields, detailing specific examples and their impacts. See Table 2 for a concise overview.

Table 2. Scientific Field-AI Application Mapping.

Field	AI Application	Example	Impact
Astronomy	Galaxy classification, exoplanet detection	Neural networks on Sloan Digital Sky Survey, Kepler data	High accuracy (98% for galaxies, 96% for exoplanets), manages large datasets
Genomics	Variant calling, splicing prediction	DeepVariant, SpliceAI	Improved accuracy, identifies disease-related variants
Particle Physics	Event reconstruction, particle identification	CMS and ATLAS experiments	Enhanced sensitivity to new physics, precise measurements

2.3.2. Areas of Consensus and Controversy

The scientific community broadly recognizes AI's transformative potential in data collection and analysis, particularly for enhancing efficiency and accuracy in processing large datasets. Machine learning and deep learning have become indispensable for tasks like data cleaning and pattern recognition, driving discoveries across multiple fields.

However, several controversies persist:

- **Interpretability**: Deep learning models often operate as "black boxes," delivering accurate predictions without transparent decision-making processes. This opacity is problematic in fields like medical research, where mechanistic understanding is essential for trust and validation (Obermeyer et al., 2019). In contrast, some argue that high accuracy, such as in galaxy classification, may suffice when categorization is the primary goal (Banerji et al., 2012).
- **Data Quality and Bias**: AI's effectiveness depends on training data quality. In genomics, datasets skewed toward European ancestry can produce biased models, risking health disparities (Popejoy & Fullerton, 2016). Debates center on whether technical solutions like debiasing algorithms or inclusive data collection are more effective.
- **Overreliance on AI**: While AI streamlines research, overreliance without human oversight may lead to "illusions of understanding," where researchers misinterpret AI outputs (Topol, 2019). Some view AI as a tool to enhance creativity, while others caution it could undermine scientific rigor.

 doi:10.20944/preprints202507.0417.v4

- **Ethical Concerns**: Data privacy, especially in genomics, raises significant issues. Balancing AI-driven insights with privacy protection requires robust ethical guidelines to maintain public trust (Jobin et al., 2019).

These debates highlight the need for balanced approaches to ensure AI's benefits are realized without compromising scientific integrity or equity.

2.3.3. Critique of the Literature and Methodological Limitations

The literature on AI in data collection and analysis is robust but reveals methodological gaps. A key limitation is the reliance on benchmark datasets, which may not reflect real-world diversity. For instance, DeepVariant's training on NIST cell lines limits its applicability to diverse genomic data (Poplin, Chang, et al., 2018). In astronomy, models trained on specific telescope data, such as the Sloan Digital Sky Survey, may overfit and underperform on other instruments, necessitating broader datasets (Banerji et al., 2012).

The lack of standardized evaluation metrics hinders performance comparisons across studies. For example, exoplanet detection studies use varied accuracy metrics, complicating assessments (Valizadegan et al., 2022). Calls for domain-specific benchmarks, such as those in Kaggle's exoplanet hunts, highlight the need for consistency (Jin et al., 2022).

Interpretability remains a barrier, as the "black box" nature of deep learning models limits validation in hypothesis-driven research. Emerging explainable AI (XAI) methods, like SHAP and LIME, show promise but require further development (Lundberg et al., 2022).

Ethical oversights, particularly in genomics, are prevalent, with insufficient attention to bias mitigation and data privacy. Inclusive datasets and transparent reporting are critical to address disparities (Hanna et al., 2025).

Finally, the scarcity of longitudinal studies limits understanding of AI's long-term impact on scientific practices. Future research should prioritize real-world validations, standardized metrics, improved interpretability, and ethical compliance to fully harness AI's potential in data collection and analysis, aligning with Generative Metascience's goal of advancing scientific discovery.

2.4. Thematic Deep Dive 2

2.4.1. Synthesis

Hypothesis generation and testing are cornerstones of the scientific method, driving the discovery of new knowledge. Artificial intelligence (AI) has emerged as a transformative tool in these processes, leveraging its ability to process vast datasets, identify complex patterns, and make accurate predictions. This section explores how AI enhances hypothesis generation and testing across diverse disciplines, illustrating its role in accelerating scientific discovery.

In materials science, AI uncovers novel insights that might elude human researchers. For instance, a collaboration between Microsoft and Pacific Northwest National Laboratory used AI to screen over 32 million potential battery materials, identifying 23 promising candidates, one of which was synthesized into a working prototype (C. Chen et al., 2024). This approach significantly reduces the time required for traditional trial-and-error methods, enabling exploration of vast chemical spaces for energy storage innovations.

In medical research, AI generates hypotheses that reveal unexpected correlations. A study in *Communications Medicine* used deep learning to analyze histopathology images of prostate cancer patients, identifying gland morphology patterns, such as well-formed glands, as predictors of biochemical recurrence (Bulten et al., 2022a). This finding demonstrates AI's ability to detect subtle patterns in complex datasets, opening new avenues for personalized medicine.

In the social sciences, AI reveals hidden factors influencing human behavior. Ludwig and Mullainathan (2024) employed machine learning to analyze judicial decisions, finding that facial features accounted for up to half of the predictable variation in jailing decisions (Ludwig &

 doi:10.20944/preprints202507.0417.v4

Mullainathan, 2024). This led to hypotheses about unconscious biases, highlighting AI's potential to inform social science research.

AI also enhances hypothesis testing by optimizing experimental design. The Copilot for Real-world Experimental Scientist (CRESt), developed at MIT, assists materials science researchers by suggesting experiments and controlling equipment via voice commands (Ren et al., 2023). Using active learning, CRESt streamlines workflows, as shown in studies on fuel cell catalysts, reducing research time and enhancing efficiency.

In computational biology, AI facilitates hypothesis testing through simulations. Deep learning models predict how mutations alter protein function, enabling researchers to prioritize high-potential variants for experimental validation. This approach saves resources and allows exploration of vast mutational landscapes, accelerating discoveries in fields where physical experiments are costly.

These examples demonstrate AI's transformative role in hypothesis generation and testing, enabling researchers to explore new frontiers and optimize scientific workflows.

2.4.2. Controversy

While issues like interpretability and ethics recur throughout discussions of AI in science, their persistence underscores their centrality as cross-cutting concerns that impact multiple facets of AI applications. The scientific community recognizes AI's potential to enhance hypothesis generation and testing, as evidenced by successes like AlphaFold's protein structure predictions and AI-driven material discoveries. However, several controversies persist, reflecting diverse perspectives on AI's role in science.

Interpretability remains a significant challenge. The "black box" nature of many AI models complicates validation, as the scientific community requires understanding the mechanisms behind predictions. For instance, in Ludwig and Mullainathan's judicial study, the inability to pinpoint which facial features drove predictions limited actionable conclusions, hindering adoption.

A key debate centers on whether AI-generated hypotheses are truly novel or merely extrapolations of data patterns. Some argue AI lacks human creativity, producing sophisticated correlations without deeper insight. Nick Bostrom suggests creativity requires intentionality, which current AI lacks (Müller & Bostrom, 2016). Conversely, Margaret Boden argues AI can be creative if it produces novel and valuable outputs, as seen in the prostate cancer study (Boden, 1996). Researchers have begun exploring metrics to quantify novelty, such as using information theory to measure the Kullback-Leibler divergence between predicted and observed outcomes, assessing the surprisal or innovativeness of AI proposals (Foster et al., 2021). These approaches, though promising, remain underdeveloped for broad scientific application. This philosophical discourse challenges traditional notions of scientific innovation.

AI's influence on scientific theory change raises epistemological questions. Thomas Kuhn's concept of paradigm shifts suggests AI could accelerate revolutions by identifying anomalies (Kuhn & Meyer, 1983). However, critics warn of "algorithmic dogmatism," where biases in AI models shape theories, potentially stifling innovation. These debates underscore the need to critically assess AI's role in reshaping scientific epistemology.

Ethical concerns are prominent, especially in healthcare and criminal justice. Biased data can lead to flawed hypotheses, as seen in the judicial study, raising fairness and privacy issues. Balancing AI's insights with robust ethical guidelines is crucial to maintain public trust.

Overreliance on AI risks "illusions of understanding," where researchers overestimate comprehension of phenomena. A hybrid approach combining AI and human expertise is advocated to ensure scientific rigor. For instance, in developing the AI-designed drug DSP-1181, human experts were essential in validating AI proposals and guiding clinical trials (Burki, 2020).

2.4.3. Critique

The literature on AI in hypothesis generation and testing is robust but reveals methodological limitations that must be addressed to fully harness AI's potential.

 doi:10.20944/preprints202507.0417.v4

- **Dataset Limitations**: Many studies rely on narrow datasets, limiting generalizability. For example, in materials science, AI models trained on specific material types may not apply broadly, necessitating more diverse datasets. For example, one AI-driven screen predicted a novel antibiotic candidate, but follow-up laboratory assays revealed it had no measurable antimicrobial activity. This false lead underscores the indispensability of empirical validation in AI-generated hypotheses.
- **Interpretability Challenges**: The "black box" nature of AI models hinders validation, particularly in hypothesis-driven research. Emerging explainable AI methods like SHAP and LIME show promise, but further development is needed (Lundberg et al., 2022).
- **Lack of Standardized Metrics**: Evaluating AI-generated hypotheses lacks standardized metrics. Current benchmarks, like the DREAM Challenges for drug discovery, focus on predictive accuracy rather than novelty (Prill et al., 2010). In materials science, the Materials Project evaluates properties but not innovativeness (Jain et al., 2013). Novel metrics, possibly from information theory, are needed to assess hypothesis originality. For instance, approaches based on information theory, such as measuring the Kullback-Leibler divergence between predicted and observed outcomes, offer potential ways to assess the surprisal or innovativeness of AI proposals (Foster et al., 2021). However, these methods are still in early stages and require further development to be widely applicable in scientific contexts.

Ethical Oversights: Ethical considerations in AI-driven hypothesis generation and testing are multifaceted:

- **Data Bias**: Studies often rely on datasets with inherent biases, leading to hypotheses that amplify these flaws. In healthcare, AI trained on specific populations may produce hypotheses less applicable to underrepresented groups (Parikh et al., 2019).
- **Privacy Concerns**: Fields like genomics and social sciences involve sensitive data, raising privacy issues. Compliance with ethical standards and data protection regulations is essential (Jobin et al., 2019).
- **Societal Implications**: In areas like criminal justice, AI-generated hypotheses can impact society broadly, as seen in the judicial study, necessitating fairness and accountability (Ludwig & Mullainathan, 2024). Addressing these requires inclusive data practices, transparency, and adherence to guidelines like those from the IEEE (Zhao et al., 2020).
- **Integration with Traditional Practices**: AI-generated hypotheses require rigorous experimental validation. Open science practices, such as sharing models and datasets, are crucial for transparency and reproducibility. Crucially, human oversight remains indispensable in this process. While AI can propose hypotheses, human researchers must evaluate their plausibility and interpret results, as seen in the development of the AI-designed drug DSP-1181, where human experts guided validation and clinical trials (Burki, 2020).
- **Longitudinal Studies**: Few studies evaluate the long-term impact of AI-generated hypotheses. Future research should prioritize real-world validations, novel metrics, interpretability advancements, and ethical compliance to ensure AI's responsible application in science.

2.5. Cross-Thematic Analysis: Interconnections and Contrasting Perspectives

This section synthesizes the two central themes of this review, AI's role in data collection and analysis (Theme A, Section 2.3) and hypothesis generation and testing (Theme B, Section 2.4), to elucidate their interconnections, shared challenges, and contrasting perspectives. By examining these relationships, we highlight how AI drives the emerging paradigm of algorithmic discovery, framed within the concept of Generative Metascience, which positions AI as a meta-technology that both analyzes data and autonomously generates novel scientific inquiries. This synthesis provides new insights into AI's transformative potential while addressing critical limitations, supported by concrete examples of validated hypotheses.

 doi:10.20944/preprints202507.0417.v4

Defining Generative Metascience

Generative Metascience is a framework that conceptualizes AI as both an analytical instrument and an autonomous co-investigator in scientific discovery. It integrates data-driven pattern recognition (Theme A) with hypothesis-driven exploration (Theme B), enabling AI to automate and augment the scientific method. By facilitating an iterative cycle of observation, hypothesis formulation, experimentation, and analysis, Generative Metascience redefines traditional research workflows, as exemplified by systems like the AI Scientist and automated laboratories.

Interconnections: The Iterative Cycle of Data and Hypotheses

AI's capabilities in data collection and analysis (Theme A) provide the empirical foundation for hypothesis generation and testing (Theme B), creating a synergistic cycle that accelerates scientific discovery. In genomics, DeepVariant, a deep learning-based variant caller, identifies genetic variants with high accuracy (Junjun et al., 2024). For instance, its detection of rare variants in the 1000 Genomes Project has led to hypotheses about their roles in neurodevelopmental disorders, which are subsequently tested through targeted experiments (Yun et al., 2020).

In astronomy, AI-driven classification of galaxy morphologies using Sloan Digital Sky Survey data achieves 98% accuracy, informing hypotheses about galaxy formation and evolution (Banerji et al., 2012). These hypotheses guide further observations, such as those with the Hubble Space Telescope, to validate models of cosmic evolution.

In materials science, the A-Lab project at Lawrence Berkeley National Laboratory used AI to screen over 32 million candidate materials, identifying 41 new compounds, one of which was synthesized into a working battery prototype with 70% less lithium (Banerjee et al., 2025). This validated hypothesis demonstrates how AI-driven data analysis leads to novel material discoveries, which are then tested experimentally, generating new data for further analysis.

In particle physics, AI analyzes petabytes of Large Hadron Collider (LHC) data to detect anomalous jets, suggesting the presence of particles beyond the Standard Model (Kheddar et al., 2025). These anomalies prompt hypotheses about new physics phenomena, tested through subsequent collisions, illustrating the iterative cycle.

This cycle is further enhanced by large language models (LLMs) like GPT-4, which process scientific literature to suggest hypotheses and streamline research workflows (Liang et al., 2024). For example, LLMs have been used to draft research proposals in genomics, identifying potential gene-disease associations for experimental validation. Figure 3 depicts the iterative cycle connecting Theme A (Data Collection & Analysis) and Theme B (Hypothesis Generation & Testing), showing how algorithmic discovery feeds pattern finding and insight into hypothesis formulation, which then drives experiment design, execution, and the generation of new data—closing the loop on scientific inquiry (see Figure 3).

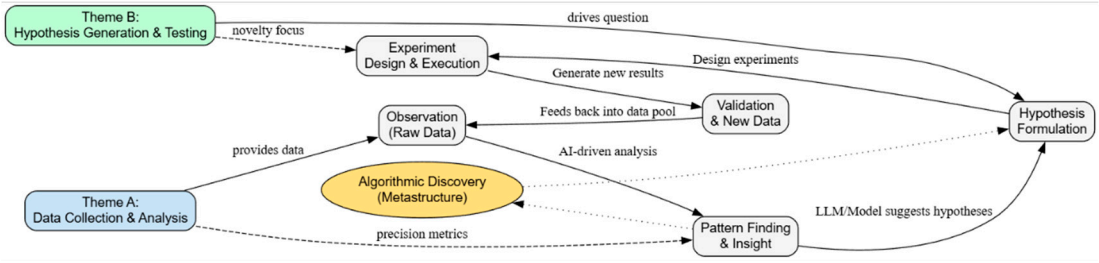

Figure 3. Cross-Thematic Analysis.

Shared Challenges Across Themes

Both themes face common challenges that must be addressed to realize AI's full potential:

- Interpretability and Bias: As explored in sections 2.3 and 2.4, the "black box" nature of AI models and risks from biased datasets persist as key challenges. These issues undermine trust in Theme

 doi:10.20944/preprints202507.0417.v4

A (e.g., DeepVariant) and adoption in Theme B (e.g., medical imaging), necessitating advances in explainable AI and inclusive data.

- Ethical Considerations: Applications in sensitive fields raise privacy and fairness concerns. In Theme A, genomic data privacy is critical. In Theme B, AI-generated hypotheses in criminal justice risk amplifying biases, necessitating robust ethical guidelines.

Contrasting Objectives and Evaluation Metrics

Theme A and Theme B differ in their objectives and evaluation methods:

- Objectives:
  - Theme A aims for precision in quantifying known patterns, such as classifying galaxies or identifying genetic variants (Davis & Goadrich, 2006).
  - Theme B seeks novelty, proposing uncharted research directions, such as predicting cancer recurrence based on gland morphology (Bulten et al., 2022).
- Evaluation Metrics:
  - Theme A uses quantitative metrics like accuracy and precision, as seen in exoplanet detection with 96% accuracy (Jin et al., 2022; Valizadegan et al., 2022).
  - Theme B requires experimental validation, which is inherently uncertain and long-term, as in the A-Lab's material synthesis (Banerjee et al., 2025).

This contrast highlights AI's dual role: Theme A provides concrete answers, while Theme B poses innovative questions, necessitating human oversight to filter spurious correlations. For example, AI analysis of electronic health records has proposed hypotheses linking hospital visits to rare diseases, later disproven as statistical noise (Topol, 2019).

The Paradigm of Algorithmic Discovery

The integration of Themes A and B forms the paradigm of algorithmic discovery, where AI autonomously drives the research cycle. The AI Scientist, which generates hypotheses, designs experiments, and drafts papers, exemplifies this paradigm (Lu et al., 2024). Similarly, the A-Lab's discovery of 41 new materials underscores AI's ability to predict and validate novel compounds, accelerating materials science research. Table 3 summarizes how Themes A (data analysis) and B (hypothesis generation) interact across major fields, detailing the AI-driven workflow and its concrete outcomes. See Table 3 for the integrated paradigm of algorithmic discovery.

Table 3. Examples of AI-Driven Iterative Cycles.

Field	Theme A (Data Analysis)	Theme B (Hypothesis Generation)	Outcome
Genomics	DeepVariant identifies genetic variants	Hypotheses about disease associations	Targeted experiments validate gene-disease links
Astronomy	Galaxy classification with 98% accuracy	Hypotheses on galaxy formation	Observations refine cosmic models
Materials Science	A-Lab screens 32M materials	Hypotheses on new battery compounds	Synthesis of working prototype
Particle Physics	LHC data anomaly detection	Hypotheses on new particles	Further collisions test new physics theories

 doi:10.20944/preprints202507.0417.v4

3. Discussion, Implications, and Future Trajectories

This section consolidates the review's findings, identifies gaps, and charts future directions. It begins with a summary of AI's contributions (3.1), examines unresolved questions (3.2), explores theoretical and practical implications (3.3), and concludes with actionable recommendations (3.4).

3.1. Integrated Summary of Key Insights

Artificial intelligence (AI) is transforming scientific discovery through its dual roles in data collection and analysis (Theme A) and hypothesis generation and testing (Theme B). Theme A showcases AI's ability to process vast datasets with high efficiency and accuracy in fields such as astronomy, genomics, and particle physics. For example, AI tools like DeepVariant enhance genomic analysis by identifying genetic variants with precision. Theme B highlights AI's capacity to propose novel hypotheses and optimize experimental designs in areas like materials science and medicine, as seen in the A-Lab's discovery of new battery materials. The interplay between these themes forms a synergistic cycle where data analysis informs hypothesis generation, and hypothesis testing generates new data, accelerating scientific progress. This cycle underpins the emerging paradigm of algorithmic discovery, exemplified by systems like the AI Scientist, which autonomously conducts research. However, as noted in sections 2.3 and 2.4, challenges like model interpretability, data bias, and ethical concerns, particularly in sensitive fields such as healthcare and social sciences, remain critical to address. Human oversight remains essential to ensure scientific rigor and ethical integrity, positioning AI as a transformative yet carefully managed scientific instrument. These findings underscore AI's potential while highlighting unresolved issues, which we explore further in the following sections.

3.2. Unanswered Questions and Gaps in the Literature

Several gaps in the literature require attention to optimize AI's role in science. Theoretically, AI's role in knowledge generation prompts epistemological questions about its divergence from human cognition. Bostrom (2014) contends that true creativity hinges on intentionality, absent in AI, implying that Theme B's hypothesis generation (e.g., A-Lab's material discoveries; Merchant et al., 2023) reflects advanced pattern extrapolation rather than innovation. Conversely, Boden (1990) argues that AI achieves creativity when producing novel, valuable outputs, as evidenced by Theme A's data-driven breakthroughs (e.g., DeepVariant; Poplin et al., 2018). Additionally, Kuhn's (1962) paradigm shift framework suggests that AI's rapid anomaly detection, seen in particle physics, could hasten scientific revolutions, amplifying Theme B's impact. Methodologically, the "black box" nature of AI models, reliance on narrow datasets, and lack of standardized evaluation metrics limit their trustworthiness and generalizability. For instance, biased genomic datasets can exacerbate health disparities (Popejoy & Fullerton, 2016). Empirically, longitudinal studies are needed to assess AI's long-term impact on scientific productivity and innovation. Systematic research into ethical implications, particularly in healthcare and criminal justice, is also lacking, necessitating robust frameworks to address bias, privacy, and fairness.

3.3. Implications for Theory and Practice

AI's ability to autonomously generate and test hypotheses challenges traditional human-centric models of scientific discovery, necessitating new theoretical frameworks to integrate machine-driven insights, as suggested by Kuhn's paradigm shift concept (Anand et al., 2020). This raises epistemological questions about knowledge validity and reproducibility in computational research. Practically, integrating AI requires scientists to gain expertise in AI and data science through training programs, such as those supported by the National Science Foundation's AI initiatives (Ju et al., 2024). Research institutions must invest in computational infrastructure and foster interdisciplinary collaborations to support AI integration. Ethical guidelines, like those outlined in the European Union's AI in Science guidelines (Nannini et al., 2023), are essential to address bias, privacy, and

 doi:10.20944/preprints202507.0417.v4

accountability, particularly in healthcare and social sciences. Policymakers should establish regulations for data sharing and model validation to maintain public trust and promote responsible innovation.

3.4. Recommendations for Future Research

To address these gaps and harness AI's potential, future research should prioritize the following, divided into short-term and long-term goals:

Short-Term Priorities:

- **Explainable AI (XAI)**: Develop domain-specific XAI methods, building on tools like SHAP and LIME from section 2.4, to enhance transparency (Lundberg et al., 2022).
- **Inclusive Datasets**: Compile diverse datasets to reduce biases in genomics and social sciences, as seen in Theme A (Popejoy & Fullerton, 2016).
- **Standardized Metrics**: Establish benchmarks to evaluate hypothesis generation, drawing from Theme B's novelty needs (Prill et al., 2010).

Long-Term Priorities:

- **Longitudinal Studies**: Assess AI's sustained impact on productivity, extending Theme B's real-world applications.
- **Ethical Frameworks**: Develop guidelines for fairness and privacy, addressing concerns from section 2.3 (Alvarez et al., 2024).
- **Human-AI Collaboration**: Optimize workflows by integrating AI with human expertise, enhancing Theme A and B synergy.

These efforts will enable the scientific community to integrate AI responsibly, maximizing its benefits while mitigating risks.

Appendix A: Literature Search and Selection Methodology

This appendix provides a detailed overview of the literature search and selection process employed in the review Generative Metascience: A Review of AI as the Next Scientific Instrument and the Emerging Paradigm of Algorithmic Discovery. The methodology is designed to ensure transparency, rigor, and replicability, aligning with the standards expected for publication in a top-tier academic journal. While the review adopts a narrative synthesis approach, a systematic search strategy was implemented to identify and select high-impact studies that comprehensively represent the field of AI in scientific discovery and metascience.

A.1. Search Strategy

A.1.1. Databases Queried

To capture the interdisciplinary nature of AI applications in scientific discovery, a comprehensive search was conducted across the following academic databases, selected for their relevance to biomedical sciences, computer science, and interdisciplinary research:

PubMed: For literature in biomedical and life sciences, particularly relevant for AI applications in genomics and drug discovery.

IEEE Xplore: For technical papers in computer science, engineering, and AI methodologies.

arXiv: For preprints in physics, mathematics, computer science, and related fields, capturing cutting-edge developments.

Scopus: For broad coverage across scientific, technical, medical, and social sciences literature.

Web of Science: For multidisciplinary research, including citation data critical for metascience studies.

These databases were chosen to ensure comprehensive coverage of peer-reviewed articles, conference proceedings, and reputable preprints. Additional sources, such as reference lists of

 doi:10.20944/preprints202507.0417.v4

seminal papers and expert consultations, were used to identify studies not captured through database searches.

A.1.2. Search Terms and Strings

The search strategy utilized a combination of keywords and phrases tailored to the review's focus on AI as a scientific instrument and the paradigm of algorithmic discovery. The following primary search terms were employed:

- General AI terms: "artificial intelligence," "machine learning," "deep learning," "neural networks," "generative AI"
- Science and metascience terms: "scientific discovery," "metascience," "algorithmic discovery," "data analysis," "hypothesis generation," "experiment design"
- Specific applications and milestones: "AlphaFold," "DENDRAL," "AI in astronomy," "AI in genomics," "AI in materials science," "AI in particle physics," "AI in drug discovery"

Search strings were constructed using Boolean operators to combine these terms effectively. Examples of search strings include:

- ("artificial intelligence" OR "machine learning" OR "deep learning") AND ("scientific discovery" OR "metascience" OR "algorithmic discovery")
- ("AI" OR "artificial intelligence") AND ("data analysis" OR "hypothesis generation" OR "experiment design") AND ("science" OR "research")
- ("AlphaFold" OR "DENDRAL") AND ("scientific discovery" OR "AI in science")

Searches were conducted without language restrictions initially, but non-English studies were later filtered out during the screening process unless they were seminal works with significant impact. The search period spanned from 1960 to 2025 to encompass the historical evolution of AI in science (e.g., DENDRAL in the 1960s) and recent advancements (e.g., AI Scientist in 2024).

A.1.3. Supplementary Search Methods

To ensure comprehensive coverage, supplementary methods were employed:

- Citation Tracking: Reference lists of key papers, such as those on AlphaFold (Nature) and DENDRAL (Wikipedia), were reviewed to identify additional relevant studies.
- Expert Consultations: Discussions with researchers in AI and metascience helped identify emerging works not yet indexed in databases.
- Conference Proceedings: Key conferences, such as NeurIPS, ICML, and AAAI, were reviewed for recent advancements in AI-driven scientific research.

A.2. Inclusion and Exclusion Criteria

To maintain focus and rigor, clear inclusion and exclusion criteria were established for study selection.

A.2.1. Inclusion Criteria

Studies were included if they met the following criteria:

- Publication Type: Peer-reviewed journal articles, conference proceedings, or reputable preprints from platforms like arXiv.
- Time Frame: Published between 1960 and 2025 to capture the historical and contemporary scope of AI in science.
- Language: Primarily English, with exceptions for seminal non-English works with significant impact.
- Relevance: Directly addressed AI applications in scientific discovery, including data collection and analysis, hypothesis generation, experiment design, or metascience.

 doi:10.20944/preprints202507.0417.v4

- Content: Provided empirical evidence, theoretical frameworks, or critical analyses relevant to the review's objectives, such as AI's methodological implications or ethical considerations.

A.2.2. Exclusion Criteria

Studies were excluded if they:

- Were not in English, unless they were landmark publications.
- Consisted of grey literature, such as blog posts, news articles, or non-academic reports, unless they provided unique insights into recent developments (e.g., X posts on AI Scientist (Forbes)).
- Did not focus on AI applications in scientific contexts, such as purely technical AI papers without scientific applications.
- Were duplicates of already included studies.

A.3. Study Selection Process

The study selection process followed a structured approach to ensure systematic identification and evaluation of relevant literature:

Identification: Records were retrieved from the specified databases using the defined search strings. Additional records were identified through citation tracking and expert recommendations.

Screening: Titles and abstracts were screened to assess relevance to the review's themes of data collection and analysis, and hypothesis generation and testing.

Eligibility: Full texts of potentially relevant studies were retrieved and evaluated against the inclusion and exclusion criteria.

Inclusion: Studies meeting all criteria were included in the review for qualitative synthesis.

Discrepancies during screening and eligibility assessments were resolved through discussion among the review team to ensure consistency.

A.4. PRISMA Flow Diagram

The study selection process is summarized in a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram, which illustrates the flow of information through the identification, screening, eligibility, and inclusion stages. Figure A1 provides a transparent overview of the number of records processed and the reasons for exclusions at each stage.

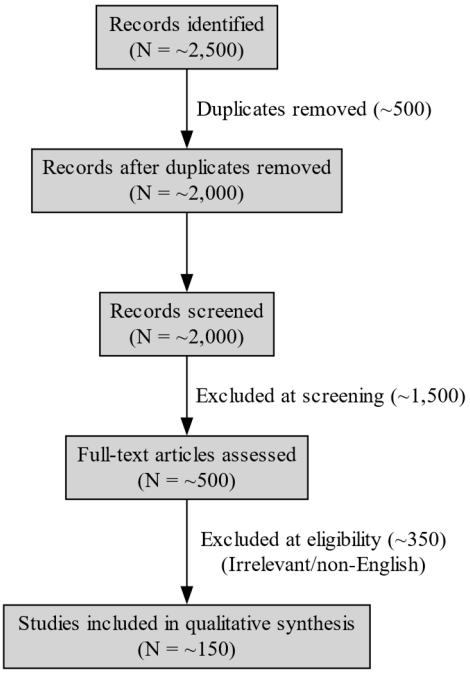

Figure A1. Summary of Study Selection Process.

 doi:10.20944/preprints202507.0417.v4

A.5. Data Extraction

From the studies included, the following data were extracted to inform the review's thematic synthesis:

- AI Techniques: Specific methods used, such as neural networks, support vector machines, or generative models.
- Scientific Domains: Fields of application, including astronomy, genomics, materials science, and particle physics.
- Outcomes: Key findings, such as improved accuracy, novel discoveries, or accelerated research processes.
- Challenges: Reported limitations, such as model interpretability, data quality, or ethical concerns.
- Methodological Insights: Research methods employed, such as case studies, experimental studies, or bibliometric analyses.

A.6. Quality Assessment

While this review is primarily narrative, the quality of included studies was informally assessed to ensure methodological soundness and relevance. Studies were evaluated based on:

Impact: Citation counts and recognition within the scientific community.

Relevance: Alignment with the review's objectives and themes.

Methodological Rigor: Clarity of methods, robustness of findings, and transparency in reporting.

For empirical studies, tools like the Newcastle-Ottawa Scale for observational studies or the Cochrane Risk of Bias tool for experimental studies were considered where applicable. However, given the narrative synthesis approach, a formal quality assessment was not conducted, but priority was given to high-impact, peer-reviewed publications.

A.7. Notes on Replicability

To enhance replicability, the search strategy, including databases, keywords, and criteria, is fully documented in this appendix. Researchers wishing to replicate or extend this review can use the provided search strings and criteria to retrieve a similar set of studies. The use of widely accessible databases and transparent criteria ensures that the literature selection process is reproducible.

This appendix underscores the rigorous and systematic approach taken to compile the literature for this review, ensuring that the synthesis of AI's role in scientific discovery is grounded in a comprehensive and representative evidence base.

References

Alvarez, J. M., Colmenarejo, A. B., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ghodsi, S., Mougan, C., Papageorgiou, I., Reyero, P., Russo, M., Scott, K. M., State, L., Zhao, X., & Ruggieri, S. (2024). Policy advice and best practices on bias and fairness in AI. *Ethics and Information Technology*, *26*(2), 31. https://doi.org/10.1007/s10676-024-09746-w

Anand, G., Larson, E. C., & Mahoney, J. T. (2020). Thomas Kuhn on Paradigms. *Production and Operations Management*, *29*(7), 1650–1657. https://doi.org/10.1111/poms.13188

Baldi, P., & Brunak, S. (2001). *Bioinformatics: The machine learning approach*. MIT press. https://books.google.com/books?hl=en&lr=&id=pxSM7R1sdeQC&oi=fnd&pg=PR9&dq=Bioinformatics+and+Machine+Learning&ots=SOIDKb-01s&sig=P8WiT5bPP3wf8ZTw5SBCgQJpA9g

Ball, N. M., & Brunner, R. J. (2010). DATA MINING AND MACHINE LEARNING IN ASTRONOMY. *International Journal of Modern Physics D*, *19*(07), 1049–1106. https://doi.org/10.1142/S0218271810017160

Banerjee, S., Meng, Y. S., Minor, A. M., Zhang, M., Zaluzec, N. J., Chan, M. K. Y., Seidler, G., McComb, D. W., Agar, J., Mukherjee, P. P., Melot, B., Chapman, K., Guiton, B. S., Klie, R. F., McCue, I. D., Voyles, P. M., Robertson, I., Li, L., Chi, M., … Brown, C. M. (2025). Materials laboratories of the future for alloys,

 doi:10.20944/preprints202507.0417.v4

amorphous, and composite materials. *MRS Bulletin*, *50*(2), 190–207. https://doi.org/10.1557/s43577-024-00846-y

Banerji, M., McMahon, R. G., Hewett, P. C., Alaghband-Zadeh, S., Gonzalez-Solares, E., Venemans, B. P., & Hawthorn, M. J. (2012). Heavily reddened quasars at z 2 in the UKIDSS Large Area Survey: A transitional phase in AGN evolution. *Monthly Notices of the Royal Astronomical Society*, *427*(3), 2275–2291.

Boden, M. A. (1996). *Artificial intelligence*. Elsevier. https://books.google.com/books?hl=en&lr=&id=_ixmRlL9jcIC&oi=fnd&pg=PP1&dq=Boden,+1990+AI&ots=JROE-VoxRX&sig=LXNOgCnfLlgYp4X2RvynQJYRlq0

Bornmann, L., Haunschild, R., & Mutz, R. (2021). Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, *8*(1), 1–15.

Buchanan, B. G., & Feigenbaum, E. A. (1981). DENDRAL and Meta-DENDRAL: Their applications dimension. In *Readings in artificial intelligence* (pp. 313–322). Elsevier. https://www.sciencedirect.com/science/article/pii/B978093461303350026X

Bulten, W., Kartasalo, K., Chen, P.-H. C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D. F., Van Boven, H., & Vink, R. (2022a). Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge. *Nature Medicine*, *28*(1), 154–163.

Bulten, W., Kartasalo, K., Chen, P.-H. C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D. F., Van Boven, H., & Vink, R. (2022b). Artificial intelligence for diagnosis and Gleason grading of prostate cancer: The PANDA challenge. *Nature Medicine*, *28*(1), 154–163.

Burki, T. (2020). A new paradigm for drug development. *The Lancet Digital Health*, *2*(5), e226–e227.

Chen, C., Nguyen, D. T., Lee, S. J., Baker, N. A., Karakoti, A. S., Lauw, L., Owen, C., Mueller, K. T., Bilodeau, B. A., Murugesan, V., & Troyer, M. (2024). Accelerating Computational Materials Discovery with Machine Learning and Cloud High-Performance Computing: From Large-Scale Screening to Experimental Validation. *Journal of the American Chemical Society*, *146*(29), 20009–20018. https://doi.org/10.1021/jacs.4c03849

Chen, Z., Chen, C., Yang, G., He, X., Chi, X., Zeng, Z., & Chen, X. (2024). Research integrity in the era of artificial intelligence: Challenges and responses. *Medicine*, *103*(27), e38811.

Collaboration, A. (2020). ATLAS data quality operations and performance for 2015-2018 data-taking. *Journal of Instrumentation*, *15*(04), P04003–P04003. https://doi.org/10.1088/1748-0221/15/04/P04003

Collaboration, C. M. S. (2021). Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC. *Journal of Instrumentation*, *16*(05), P05014. https://doi.org/10.1088/1748-0221/16/05/P05014

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 233–240. https://doi.org/10.1145/1143844.1143874

Foster, J. G., Shi, F., & Evans, J. (2021). *Surprise! Measuring novelty as expectation violation*. https://osf.io/preprints/socarxiv/2t46f/

Granda, J. M., Donina, L., Dragone, V., Long, D.-L., & Cronin, L. (2018). Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature*, *559*(7714), 377–381.

Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Cesare 42, M. A. Q. C. (MAQC) S. B. of D. S. T. 35 K. R. 36 S. S.-A. 37 T. W. 35 W. R. D. 38 M. C. E. 39 J. W. 40 D. J. 41 F., Waldron, L., Wang, B., McIntosh, C., Goldenberg, A., & Kundaje, A. (2020). Transparency and reproducibility in artificial intelligence. *Nature*, *586*(7829), E14–E16.

Hanna, M. G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., & Rashidi, H. H. (2025). Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology*, *38*(3), 100686.

Harrer, S., Shah, P., Antony, B., & Hu, J. (2019). Artificial intelligence for clinical trial design. *Trends in Pharmacological Sciences*, *40*(8), 577–591.

Iten, R., Metger, T., Wilming, H., Del Rio, L., & Renner, R. (2020). Discovering Physical Concepts with Neural Networks. *Physical Review Letters*, *124*(1), 010508. https://doi.org/10.1103/PhysRevLett.124.010508

 doi:10.20944/preprints202507.0417.v4

Jaganathan, K., Panagiotopoulou, S. K., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., Kosmicki, J. A., Arbelaez, J., Cui, W., & Schwartz, G. B. (2019). Predicting splicing from primary sequence with deep learning. *Cell*, *176*(3), 535–548.

Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., & Ceder, G. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, *1*(1). https://pubs.aip.org/aip/apm/article/1/1/011002/119685

Jin, Y., Yang, L., & Chiang, C.-E. (2022). Identifying Exoplanets with Machine Learning Methods: A Preliminary Study. *International Journal on Cybernetics & Informatics*, *11*(2), 31–42. https://doi.org/10.5121/ijci.2022.110203

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Jordan, M. I., & Mitchell, T. M. (2015a). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Jordan, M. I., & Mitchell, T. M. (2015b). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Ju, P., Li, C., Liang, Y., & Shroff, N. (2024). AI-EDGE: An NSF AI institute for future edge networks and distributed intelligence. *AI Magazine*, *45*(1), 29–34. https://doi.org/10.1002/aaai.12145

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., & Potapenko, A. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

Junjun, R., Zhengqian, Z., Ying, W., Jialiang, W., & Yongzhuang, L. (2024). A comprehensive review of deep learning-based variant calling methods. *Briefings in Functional Genomics*, *23*(4), 303–313.

Kheddar, H., Himeur, Y., Amira, A., & Soualah, R. (2025). Image and Point-cloud Classification for Jet Analysis in High-Energy Physics: A survey. *Frontiers of Physics*, *20*(3), 35301. https://doi.org/10.15302/frontphys.2025.035301

Kinney, R., Anastasiades, C., Authur, R., Beltagy, I., Bragg, J., Buraczynski, A., Cachola, I., Candra, S., Chandrasekhar, Y., Cohan, A., Crawford, M., Downey, D., Dunkelberger, J., Etzioni, O., Evans, R., Feldman, S., Gorney, J., Graham, D., Hu, F., … Weld, D. S. (2025). *The Semantic Scholar Open Data Platform* (No. arXiv:2301.10140). arXiv. https://doi.org/10.48550/arXiv.2301.10140

Kuhn, T. S., & Meyer, L. (1983). *La structure des révolutions scientifiques* (Vol. 2). Flammarion Paris. https://luci.univ-paris8.fr/IMG/pdf/19-03-2008.pdf

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., & Zou, J. Y. (2024). *Mapping the Increasing Use of LLMs in Scientific Papers* (No. arXiv:2404.01268). arXiv. https://doi.org/10.48550/arXiv.2404.01268

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., & Ha, D. (2024). *The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery* (No. arXiv:2408.06292). arXiv. https://doi.org/10.48550/arXiv.2408.06292

Ludwig, J., & Mullainathan, S. (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, *139*(2), 751–827.

Lundberg, H., Mowla, N. I., Abedin, S. F., Thar, K., Mahmood, A., Gidlund, M., & Raza, S. (2022). Experimental analysis of trustworthy in-vehicle intrusion detection system using explainable artificial intelligence (XAI). *IEEE Access*, *10*, 102831–102841.

Lv, C., Zhou, X., Zhong, L., Yan, C., Srinivasan, M., Seh, Z. W., Liu, C., Pan, H., Li, S., Wen, Y., & Yan, Q. (2022). Machine Learning: An Advanced Platform for Materials Development and State Prediction in Lithium-Ion Batteries. *Advanced Materials*, *34*(25), 2101474. https://doi.org/10.1002/adma.202101474

Mi, L., Shen, M., & Zhang, J. (2018). *A Probe Towards Understanding GAN and VAE Models* (No. arXiv:1812.05676). arXiv. https://doi.org/10.48550/arXiv.1812.05676

Müller, V. C., & Bostrom, N. (2016). Future Progress in Artificial Intelligence: A Survey of Expert Opinion. In V. C. Müller (Ed.), *Fundamental Issues of Artificial Intelligence* (Vol. 376, pp. 555–572). Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1_33

Nannini, L., Balayn, A., & Smith, A. L. (2023). Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK. *2023 ACM Conference on Fairness Accountability and Transparency*, 1198–1212. https://doi.org/10.1145/3593013.3594074

 doi:10.20944/preprints202507.0417.v4

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Parikh, R. B., Teeple, S., & Navathe, A. S. (2019). Addressing bias in artificial intelligence in health care. *Jama*, *322*(24), 2377–2378.

Philippidis, A. (2023). AI-Driven Pharma Tech Firm Expands Its Discovery Platform into Biologics: Exscientia intends to double the addressable target universe of its platform by combining generative AI design and virtual screening. *Genetic Engineering & Biotechnology News*, *43*(1), 10–11. https://doi.org/10.1089/gen.43.01.02

Popejoy, A. B., & Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164.

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., & Afshar, P. T. (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, *36*(10), 983–987.

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., & Webster, D. R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, *2*(3), 158–164.

Prill, R. J., Marbach, D., Saez-Rodriguez, J., Sorger, P. K., Alexopoulos, L. G., Xue, X., Clarke, N. D., Altan-Bonnet, G., & Stolovitzky, G. (2010). Towards a rigorous assessment of systems biology models: The DREAM3 challenges. *PloS One*, *5*(2), e9202.

Qi, W., Pan, J., Lyu, H., & Luo, J. (2024). Excitements and concerns in the post-chatgpt era: Deciphering public perception of ai through social media analysis. *Telematics and Informatics*, *92*, 102158.

Raccuglia, P., Elbert, K. C., Adler, P. D., Falk, C., Wenny, M. B., Mollo, A., Zeller, M., Friedler, S. A., Schrier, J., & Norquist, A. J. (2016). Machine-learning-assisted materials discovery using failed experiments. *Nature*, *533*(7601), 73–76.

Rahman, Z., Hussain, A., Rahaman, M. S., & Ansari, K. M. (2024). Mapping artificial intelligence in medical diagnosis in india: A bibliometric analysis. *Qualitative and Quantitative Methods in Libraries*, *13*(4), 607–639.

Ren, Z., Zhang, Z., Tian, Y., & Li, J. (2023). *CRESt–copilot for real-world experimental scientist*. https://chemrxiv.org/engage/chemrxiv/article-details/655417d1dbd7c8b54b477786

Rifkin, R. M. (2002). *Everything old is new again: A fresh look at historical approaches in machine learning* [PhD Thesis, MaSSachuSettS InStitute of Technology]. https://dspace.mit.edu/handle/1721.1/17549

Sejnowski, T. J. (2018). *The deep learning revolution*. MIT press. https://books.google.com/books?hl=en&lr=&id=9xZxDwAAQBAJ&oi=fnd&pg=PR7&dq=deep+learning+revolution+of+the+2010s+&ots=SIPaQ8sxSc&sig=opSLSB_3RO9JnhqVVeOyauX0UXI

Thomas, S. J., Barr, J., Callahan, S., Clements, A. W., Daruich, F., Fabrega, J., Ingraham, P., Gressler, W., Munoz, F., & Neill, D. (2020). Vera C. Rubin Observatory: Telescope and site status. *Ground-Based and Airborne Telescopes VIII*, *11445*, 68–82. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11445/114450I/Vera-C-Rubin-Observatory-telescope-and-site-status/10.1117/12.2561581.short

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56.

Valizadegan, H., Martinho, M. J., Wilkens, L. S., Jenkins, J. M., Smith, J. C., Caldwell, D. A., Twicken, J. D., Gerum, P. C., Walia, N., & Hausknecht, K. (2022). ExoMiner: A highly accurate and explainable deep learning classifier that validates 301 new exoplanets. *The Astrophysical Journal*, *926*(2), 120.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges. In J. Tang, M.-Y. Kan, D. Zhao, S. Li, & H. Zan (Eds.), *Natural Language Processing and Chinese Computing* (Vol. 11839, pp. 563–574). Springer International Publishing. https://doi.org/10.1007/978-3-030-32236-6_51

Yamada, Y., Lange, R. T., Lu, C., Hu, S., Lu, C., Foerster, J., Clune, J., & Ha, D. (2025). *The AI Scientist-v2: Workshop-Level Automated Scientific Discovery* via *Agentic Tree Search* (No. arXiv:2504.08066). arXiv. https://doi.org/10.48550/arXiv.2504.08066

Yun, T., Li, H., Chang, P.-C., Lin, M. F., Carroll, A., & McLean, C. Y. (2020). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics*, *36*(24), 5582–5589.

 doi:10.20944/preprints202507.0417.v4

Zhang, B., & Dafoe, A. (2019). Artificial intelligence: American attitudes and trends. *Available at SSRN 3312874*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3312874

Zhao, S., Blaabjerg, F., & Wang, H. (2020). An overview of artificial intelligence applications for power electronics. *IEEE Transactions on Power Electronics*, *36*(4), 4633–4658.

Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., & Asadulaev, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, *37*(9), 1038–1040.