AI-Driven Automation Can Become the Foundation of Next-Era Science of Science Research

Anonymous Author(s)
Affiliation
Address
email

Abstract

The Science of Science (SoS) explores the mechanisms underlying scientific discovery, and offers valuable insights for enhancing scientific efficiency and fostering innovation. Traditional approaches often rely on simplistic assumptions and basic statistical tools, such as linear regression and rule-based simulations, which struggle to capture the complexity and scale of modern research ecosystems. The advent of artificial intelligence (AI) presents a transformative opportunity for the next generation of SoS, enabling the automation of large-scale pattern discovery and uncovering insights previously unattainable. This paper offers a forward-looking perspective on the integration of Science of Science with AI for automated research pattern discovery and highlights key open challenges that could greatly benefit from AI. We outline the advantages of AI over traditional methods, discuss potential limitations, and propose pathways to overcome them. Additionally, we present a preliminary multi-agent system as an illustrative example to simulate research societies, showcasing AI's ability to replicate real-world research patterns and accelerate progress in Science of Science research.

1 Introduction

Science of Science (SoS), a pivotal and rapidly evolving field, serves as a strategic compass for guiding the trajectory of scientific and technological progress. By analyzing the complex dynamics of research collaboration and scientific output across geographic and temporal scales, it sheds light on the factors that drive creativity and the emergence of scientific discoveries, with the goal of developing tools and policies to accelerate scientific advancement [24]. Unlike broader social sciences that examine societal structures, SoS delves deep into the mechanisms that fuel scientific breakthroughs [9, 86, 47]—illuminating the hidden forces that propel discovery and transformation. Ultimately, SoS underscores that groundbreaking advancements are not solely the result of talented minds and quality data, but are profoundly shaped by effective resource allocation, supportive policies and well-designed organizational structures [96, 98].

In recent years, the deep fusion of AI and SoS has become more feasible and promising than ever before. First, the increasing availability of large-scale scholarly data—publications, funding records, and collaboration networks—provides unprecedented opportunities to gain deeper insights into the evolution of scientific progress. Second, rapid advancements in AI technologies, such as large language models (LLMs), along with improvements in computational power, have greatly enhanced our ability to analyze and interpret complex scientific information with unprecedented accuracy and scale. These technological breakthroughs mark a critical moment for integrating AI into SoS, paving the way for a more data-driven approach to understanding and guiding research pattern discovery. While some recent works have begun exploring autonomous scientific discovery, the field remains in its infancy, and there is still much progress to be made before realizing its full potential.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

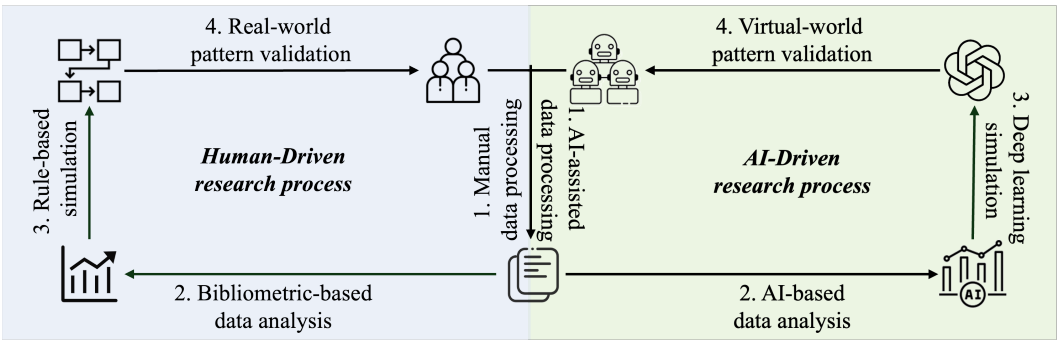

Figure 1: An illustration comparing human-driven and AI-driven research processes in the SoS, highlighting step-by-step differences across four key stages in order: *data processing*, *data analysis*, *system simulation*, and *pattern validation*.

In this paper, we take a step forward by providing the first glimpse into the integration of AI and SoS for automated research pattern discovery. **We take the position that AI has the potential to revolutionize SoS, enabling the next generation of research by not only automating traditional research processes but also providing a sandbox for SoS research, allowing scientists to observe research processes in action and validate their hypotheses.** As illustrated in Fig. 1, traditional SoS methods have primarily relied on manual data processing, bibliometric-based data analysis, rule-based system simulations, and real-world pattern validation. In contrast, AI-driven SoS leverages automated techniques to assist scientists in processing and analyzing data while offering more advanced and comprehensive systems for simulation and validation. This shift from human-driven to AI-driven methodologies unlocks the potential for more efficient, scalable, and data-driven analysis, ultimately providing deeper and more actionable insights into the mechanisms that shape scientific progress. Thus, we define AI for SoS (AI4SoS) as a cross-disciplinary field that not only focuses on facilitating each step within the research process but also aims to achieve fully automated SoS research to uncover the hidden forces driving scientific innovation. This distinguishes AI4SoS from existing AI for Science (AI4S) approaches, which focus on using AI tools to solve domain-specific scientific problems [25, 1, 13]. To better differentiate AI4SoS from AI4S, we illustrate differences in Table 1.

To consolidate our insights, we propose a forward-looking hierarchy of AI4SoS automation in Sec. 2.2. In Sec. 3, we highlight critical open problems in SoS where AI offers advantages. Despite its promise, we discuss challenges such as data bias in Sec. 4. We also propose possible pathways to overcome these challenges. Lastly, we introduce a preliminary multi-agent system to simulate research societies in Sec. 5, illustrating AI's capability to enable fully automated pattern discovery. We show related work, alternative views and impact statement in Appx. A, D, and F, respectively.

2 AI for Science of Science

2.1 Definition

AI for SoS (AI4SoS) refers to the application of AI techniques to analyze, simulate, and validate the pattern of scientific research. It aims to leverage AI to study key aspects of the scientific ecosystem, including research productivity (e.g. individual published paper count), collaboration network (e.g. interdisciplinary research collaboration), and the factors driving the advancement of scientific knowledge (e.g. funding and policy). Specifically, AI can drive the SoS research process by automatically applying methods such as machine learning, data mining, and computational simulations, thereby uncovering scientific patterns.

2.2 Hierarchy of Automation Degree in AI4SoS

The integration of AI techniques into scientific research follows a progressive hierarchy, reflecting the increasing autonomy and sophistication of AI systems in advancing the SoS field. As illustrated in Fig. 2, we define five levels of autonomy, ranging from no AI involvement in pattern recognition and analysis to full autonomy in uncovering new scientific insights and guiding research strategies.

Table 1: Comparison between AI for Science and AI for Science of Science.

Feature	AI for Science	AI for Science of Science
Focus	Solving domain-specific scientific problems.	Understanding mechanisms of scientific progress to facilitate and accelerate research.
Approach	Direct application of AI to address scientific challenges.	Meta-level analysis to enhance the research process.
Examples	Predicting weather, designing new drugs, optimizing materials.	Studying research collaboration trends, analyzing innovation triggers, mapping knowledge growth.

Level 0: Non-automated SoS Discovery At this level, scientific pattern discovery is entirely human-driven and relies on traditional statistical methods. Researchers apply fundamental techniques such as probabilistic models, linear regression, and hypothesis testing to analyze scientific data and uncover patterns. AI is not involved in the process, and all tasks are conducted manually using well-established statistical procedures. Notable studies in this domain include the application of regression analysis to identify research trends [80], correlation analysis to examine relationships between variables [5], and statistical estimation methods to explain observed scientific phenomena [59, 106].

Level 1: AI-Assisted SoS Discovery In Level 1, AI only supports scientific data processing. Specifically, AI methods are able to transform real-world scientific data into a more comprehensible form, including tasks such as completing and structuring bibliometric data, extracting key features such as author networks and institutional collaborations, and converting text information (e.g., papers, scientists) into embedding representations, thereby enhancing the efficiency and accuracy of data handling. However, AI's role remains supplementary, with human researchers still conducting data analysis, understanding and prediction. From the perspective of AI4SoS, some related works include: utilizing text-to-embedding methods for mapping papers to vector space [85], extracting key information from papers using named entity recognition [99], and constructing networks for faculty mobility [15].

Level 2: Partially Automated SoS Discovery In Level 2, AI techniques (e.g., supervised learning), play a central role in analyzing scientific data, enabling tasks such as predicting emerging trends, research hotspots and collaboration opportunities, based on historical patterns. This marks a shift from AI-assisted data processing to AI-driven data analysis. However, in this level, AI struggles to design and implement experiments automatically. For instance, a simulation environment that can automatically conduct scientific experiments is not available, therefore it is difficult to model hidden dynamic processes within the scientific ecosystem. Related works include the use of machine learning models to predict individual paper citation counts [102], neural networks for forecasting research trends and generating novel ideas [48], clustering publications based on citation relationships [92], and applying structural topic models to extract topics from scientific texts [33].

Level 3: Highly Automated SoS Discovery In Level 3, AI not only drives the analysis but also designs and implements experiments to simulate scientific patterns in the real world. In this case, researchers can compare results generated by simulation systems and those in the real world to explore strategies in SoS for potential real-world applications. While AI can support automatic experiment conduction, human supervision is required to define the specific application scenarios and corresponding experimental parameters (e.g., scientist information, boundary conditions) based on system feedback. Consequently, the authenticity and rationality of the system depends on whether the researchers have considered all relevant factors, making the automatic pattern validation difficult. Research at this level is still in its early stages, including systems simulating specific research scenarios to propose hypotheses [27], AI predicting outcomes under different simulation conditions to provide insights into collaboration patterns [90]. and systems reproducing historical events based on specific environmental settings [105].

Level 4: Fully Automated SoS Discovery Level 4, the ultimate stage, represents complete automatic discovery in SoS. An AI-based virtual research society is conducted for end-to-end SoS discovery, including pattern analysis, prediction, and validation. Compared to systems in Level 3, systems in Level 4 function with continuous AI-based feedback loops to autonomously assess research plans and results to dynamically adjust parameters such as experimental settings, enabling virtual-world pattern validation as an alternative to real-world social experiments that may be aggressive. At this stage,

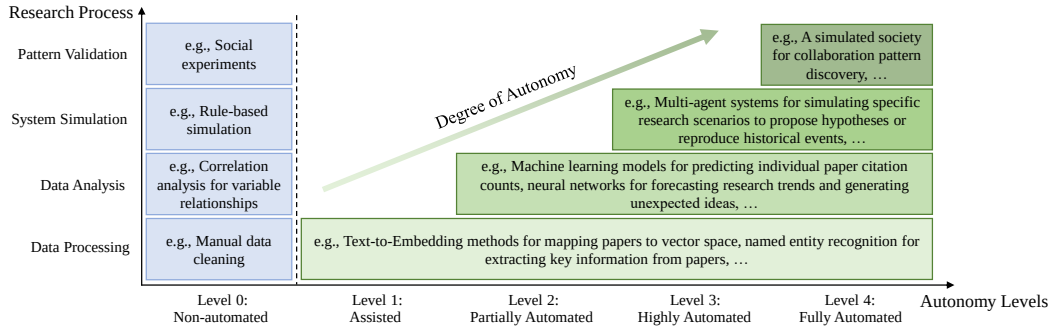

Figure 2: An overview of the five progressively advancing levels of autonomy in AI4SoS, with more green areas indicating that higher levels correspond to greater degrees of autonomy. Current research is primarily at Level 2 or below, with very limited work at Level 3, while fully automated SoS discovery remains in the prospective stage.

novel scientific insights can be discovered without human intervention, and systems can adapt to new data and incorporate new insights in real time. Ethical and governance frameworks are embedded, aligning the system's actions with established guidelines for scientific integrity and accountability.

Currently, most research remains at Level 2 or below, with limited progress observed at Level 3, while fully automated SoS discovery is still in the exploratory stage. Looking ahead, several potential tasks are envisioned, including automated discovery of new collaboration patterns within the simulated scientific community [90], systems capable of simulating and conducting experiments in real-world settings [52], and AI that continuously refines research directions based on emerging data [69].

3 Advantages of Automatic SoS Discovery

In this section, we delve into critical open problems within the SoS that stand to benefit substantially from AI-driven automation. These problems are categorized into two primary areas: *Forecasting Trends in Technology and Innovation* and *Understanding the Dynamics of Research Society*. For each subproblem, we provide a brief background and outline key opportunities where AI offers advantages.

3.1 Forecasting Trends in Technology and Innovation

3.1.1 Background of Problem

Accurately forecasting the trajectory of science and technology is a crucial aspect of SoS, as it informs decisions related to funding, policy-making, and research prioritization. Two major challenges are predicting technological trends and identifying interdisciplinary opportunities.

The Trend in Technological Development Technological development follows intricate and often non-linear trajectories, making prediction difficult. To predict these trends, it is essential to understand which technologies are gaining momentum, identify emerging breakthroughs, and anticipate when they will transition from research to real-world applications [35]. Traditional methods, such as historical data analysis, often fall short in scalability and struggle to keep pace with rapid advancements.

The Interdisciplinary Future of Innovation Interdisciplinary research, which often serves as the pivotal role for major breakthroughs, presents another significant challenge. With the rapid growth of scientific literature across diverse fields, manual identification of promising cross-disciplinary opportunities has become increasingly unfeasible [11]. The complexity and scale of this task call for automated solutions capable of discovering novel connections across fields.

3.1.2 Advantages of AI4SoS

AI offers an opportunity for tackling challenges in the SoS by leveraging its capacity to process vast datasets and identify complex patterns beyond human discernment. In the context of forecasting tech-

nological development, AI models can analyze citation networks, research metadata, and publication trends to detect emerging technological trajectories with enhanced precision [12].

Moreover, AI-driven methods excel in uncovering interdisciplinary opportunities by representing scientific knowledge as graph structures and employing advanced similarity metrics. Graph neural networks, for instance, have demonstrated the ability to model intricate relationships across scientific literature, facilitating the discovery of latent connections and novel collaborations across disparate domains [110]. This capability empowers researchers to target high-potential interdisciplinary collaborations, fostering innovation at the convergence of fields.

3.2 Understanding the Dynamics of Research Society

3.2.1 Background of Problem

The dynamics of research societies play a fundamental role in shaping scientific progress, which encompass how scientist research patterns evolve, how different team constructions influence the impact of research output, and how current research society influences scientists.

The Dynamics and Mechanics of Scientist Career The role of studying scientific careers is to provide personalized support to the academic community, thereby enhancing individual innovation capabilities, optimize team collaboration efficiency, and improving the allocation of research resources [24]. However, challenges include the highly individualized nature of career development paths, data scarcity and bias, and the complexity of external environmental factors [97].

The Dynamics and Mechanics of Research Team The composition and dynamics of scientific teams play a crucial role in improving research outcomes, with elements such as size, diversity, and collaboration patterns influencing team creativity and productivity [5, 100]. Over time, shifts in team structures and researcher mobility have reflected broader changes in the research landscape. Understanding these evolving dynamics presents challenges, as the relationships between team composition and research impact are multifaceted [101, 104].

The Dynamics and Mechanics of Research Society The organization and dynamics of research societies play a crucial role in shaping the progression and fairness of scientific endeavors. Studies have highlighted persistent inequalities in academic representation, participation, and recognition, both within and across nations [98, 56]. These disparities, influenced by systemic and structural factors, hinder the equitable generation and dissemination of knowledge. On a broader scale, imbalances in citation patterns and collaboration networks often reflect biases rooted in reputation and resources rather than research quality [28].

3.2.2 Advantages of AI4SoS

AI offers potential for understanding and improving the dynamics of research societies. By analyzing large-scale historical datasets—such as collaboration patterns, research trajectories, and external influences—AI can uncover critical factors driving individual career development. This enables personalized researcher support and helps institutions optimize talent management. Techniques such as predictive modeling have proven effective in tracking and forecasting team member mobility patterns [30].

Moreover, AI-driven agents can simulate complex team dynamics, providing insights into how various factors, such as diversity and team size, influence research productivity and innovation. Taking this a step further, AI can simulate entire scientific societies, not only uncovering hidden patterns and problems but also guiding the policymaking process by validating potential policies within the simulated environment. For instance, multi-agent systems have been employed to model team formation processes and predict collaboration outcomes under varying settings [90].

4 Challenges and Pathways

Achieving fully automated SoS discovery centers on effectively utilizing AI techniques to process scientific data. This endeavor involves addressing four key challenges: data-related issues, comprehensive system construction, robust system evaluation, and system explainability. For each of these challenges, we provide a detailed analysis along with potential pathways for resolution.

4.1 Data Issues

Challenges Data issues mainly include data imbalance across disciplines and training data bias. For the first issue, many disciplines, such as computer science and engineering, produce large volumes of well-structured data readily used by AI systems [22, 41]. However, other fields, such as social sciences or humanities, often suffer from smaller datasets, less structured data, or incomplete information, which makes it difficult for AI models to provide accurate predictions [49, 39]. This imbalance can lead to skewed results where AI predictions are disproportionately driven by well-represented fields, neglecting potentially valuable insights from underrepresented areas of research. Another issue is training data bias. When predicting reproducible patterns from data, machine learning models inevitably incorporate and perpetuate biases present in the data, often in opaque ways [58]. For example, the training data and alignment methods of LLMs (whether open-source or closed-source) are not fully disclosed [2, 18, 103], making it impossible to objectively assess their bias and fairness. Therefore, the fairness of machine learning becomes a heavily debated issue in applications ranging from the criminal justice system to hiring processes [65].

Pathway To address issues of data imbalance and biases in training data, constructing a large and diverse dataset is essential to improve data representativeness, ensuring coverage across various domains, groups, and contexts. Several large-scale, cross-disciplinary academic datasets are currently available for SoS research, including the Microsoft Academic Graph (MAG) [87], Open Academic Graph (OAG)[108], and SciSciNet [54], as summarized in Table 2. In the process of data auditing and filtering, it is crucial to examine data sources and mitigate any potential historical or socio-cultural biases to ensure the dataset is free from implicit biases [81]. Additionally, employing multi-annotator strategies, conducting group balance checks, and performing fairness evaluations can further ensure the fairness and diversity of the dataset [73]. These measures not only enhance the model's generalization ability but also reduce unfairness stemming from data biases.

4.2 Comprehensive System Construction

Challenges Simulating a research society using AI for fully automated SoS discovery, particularly through an agent-based system, presents numerous challenges. Each scientist-agent requires detailed modeling of their research expertise, career trajectory, and collaborative networks, which are often too complex to be fully captured in the simulation system [68, 26]. Critical but unobservable factors, such as internal cognitive processes and informal discussions that drive real-world decision-making, remain challenging to replicate accurately. These limitations inevitably make simulations discrete and less representative of actual societal dynamics. Moreover, the simulation process itself introduces complexities. Aligning the simulated timeline with real-world events necessitates careful calibration; for instance, determining how many simulation epochs correspond to a year in reality [43]. Determining the appropriate size of the simulated society is also crucial; an overly small-scale model risks failing to capture the emergent behaviors of a real research ecosystem, while an overly large model may become impractical to manage and analyze [82, 7]. Another pressing challenge lies in bias amplification when designing AI systems—a concern that builds on the broader implications of how AI interacts with societal structures. Since AI systems are often designed to optimize based on historical data of SoS, they risk perpetuating existing paradigms, funding trends, and citation networks. This aligns with the well-documented "rich get richer" effect in citation and funding dynamics [21, 79, 40]. If an AI system prioritizes high-impact metrics, it may inadvertently favor mainstream topics and established researchers, further marginalizing unconventional or disruptive ideas. Without explicit mechanisms to value novelty and diversity, such systems could unintentionally confine the scientific community to existing trends, hindering pathways to groundbreaking innovation. Lastly, the system must account for unexpected exceptions to ensure the simulation operates smoothly and continuously for fully automated scientific discovery. Striking a balance between realism and feasibility remains a persistent and fundamental challenge in these simulations.

Pathway Several potential pathways can help address these complexities. With the continuous advancement of LLMs' comprehensive capabilities, handling complex multi-level modeling is becoming increasingly feasible. By defining agent models with distinct roles and appropriately assigning tasks, the behaviors of scientists at various levels can be more accurately simulated [75]. Fine-tuning LLMs on extensive academic datasets can further optimize the behavioral patterns of agents [31], enhancing their adaptability to reflect real-world research dynamics. One solution for timeline alignment is to build flexible, dynamic calibration techniques that adjust the simulation's

temporal parameters based on context and event-driven data [105]. In determining the appropriate scale for the simulated society, agent-based sampling methods (random or rule-based) or dynamic population expansion techniques can be utilized [90]. When addressing bias in AI systems, it is crucial to consider the nature of SoS, a discipline dedicated to analyzing historical data and uncovering biases or patterns within the scientific community. To ensure alignment between simulations and real-world dynamics, it is essential to incorporate these biases into SoS studies, as AI designed for this field seeks to enhance and advance SoS research. At the same time, such biases can be mitigated through targeted adjustments to system parameters. For instance, to counteract the "rich get richer" effect in citations, one effective approach could involve reducing the likelihood of citing highly cited papers when an agent selects a reference. Instead, assigning higher probabilities to less-cited, more novel papers can help promote diversity in citation practices and encourage the exploration of unconventional ideas. Moreover, the system can integrate robust anomaly detection and recovery mechanisms to handle unexpected situations. Using unsupervised learning techniques (such as clustering), the model can identify deviations from expected behaviors and adjust simulation parameters accordingly to ensure stability and continuity [3]. These potential solutions try to strike a balance between realism and operational feasibility, providing a technological foundation for research society simulations.

4.3 Comprehensive System Evaluation

Challenges Evaluating the validity of outputs generated by AI systems in the field of SoS is a complex and multifaceted challenge. SoS research addresses a broad range of problems and lacks unified evaluation standards, with different tasks often necessitating tailored metrics [58]. Moreover, innovation—a key attribute of AI outputs—is inherently subjective and context-dependent, making it difficult to quantify accurately using traditional methods [90, 14]. Validity assessments also heavily rely on specific domain contexts. However, the interdisciplinary nature of SoS compounds the complexity, requiring the integration of knowledge and evaluation standards from diverse fields. Additionally, the dynamic nature and long-term implications of AI-generated outputs present further challenges, as their true impact on scientific progress often cannot be evaluated in the short term [8]. Addressing this requires advanced tools, such as time-series analysis and virtual scientist simulations, to facilitate longitudinal tracking. Furthermore, AI-generated scientific recommendations may raise ethical issues and have far-reaching consequences for scientific communities and research practices [55]. Therefore, a comprehensive and adaptable evaluation framework is necessary, integrating scientometric methodologies, multidisciplinary expert reviews, dynamic analytical approaches, and stringent ethical guidelines.

Pathway To address these challenges, appropriate solutions can be implemented. First, collaborating with domain experts to define task-specific evaluation metrics is essential, and then quantitative evaluation methods based on scientometrics should be developed. For instance, citation counts can be used as a measure of influence when evaluating the impact of system outputs, and they can also track knowledge flow [58]. In simulating a scientist's career, individual impact metrics such as the h-index, which reflects both productivity and impact, can be applied. Additionally, to assess output novelty, feasible approaches include large model-based peer-review scoring [61, 90] or calculating the Z-score for each pairing of referenced journals [14]. With the ongoing expansion of LLMs' expertise and improved reasoning capabilities, interdisciplinary testing and long-term large-scale simulations have become increasingly feasible. Moreover, LLMs are now being employed in social simulations [105], assuming role-based agents. In terms of ethical and social impacts, aligning model preferences and improving transparency can partially address ethical concerns and enhance user trust, while ethical benchmarks [64, 37] can be used to test the validity of system outputs. By integrating these strategies, a multidimensional evaluation framework can be established.

4.4 Explainability and Causal Inference

Challenges While the AI framework emphasizes automated discovery and evaluation, it lacks mechanisms to explain the causal pathways behind AI-generated outputs [32, 78]. This limitation makes it difficult for researchers and policymakers to trust and adopt AI-driven insights, as they may not fully understand the underlying logic or relationships. Moreover, the complex and interdisciplinary nature of SoS often involves interactions between numerous variables, such as collaborations, funding patterns, and citation networks [88, 24], which cannot be adequately captured through correlation-based approaches. Without explicit causal explanations, it is challenging to ensure the auditability, accountability, and interpretability of the system, undermining its credibility and ethical alignment.

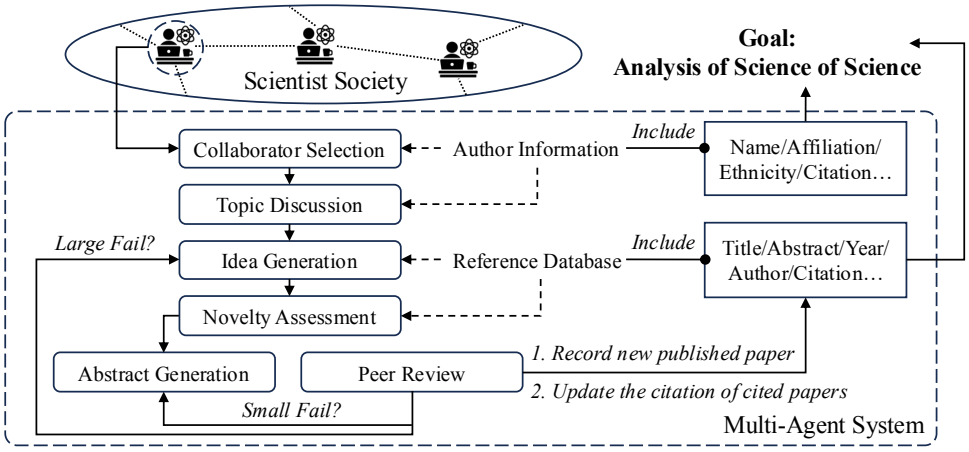

Figure 3: The overview of our preliminary multi-agent system for scientific collaboration simulation. We place the simulation within a community of scientists. After a scientist leads his/her team in submitting a paper, it undergoes peer review. If accepted, it is added to the reference database and can be cited by other scientists in subsequent epochs. Due to varying author information, the citation count of the final research output differs, then we can analyze the correlation between them—understanding the dynamics of research organizations, which is important in the field of SoS.

Pathway To address these challenges, it is crucial to introduce causal modeling [71, 23] and explainable AI (XAI) [20, 60] techniques to assist in interpreting and validating simulation results. Approaches such as Counterfactual Analysis can clarify the logical origins of AI-driven recommendations or discoveries, making the reasoning process more transparent. Relevant methods in the SoS domain include causal inference techniques like Propensity Score Matching (PSM) and Coarsened Exact Matching (CEM), which are useful for identifying causal relationships in complex systems [46, 36]. Additionally, causal graphical models and structural equation modeling (SEM) can be applied to analyze scientific impact by modeling the flow of influence across variables such as collaboration networks or funding distributions [16, 44, 50]. These tools provide a robust foundation for explaining AI-generated outputs.

5 Proof-of-Concept Studies

In this section, we present case studies to illustrate a practical application scenarios in AI4SoS. Specifically, by constructing a simplified preliminary multi-agent system to replicate phenomena observed in real-world scientific societies and uncover underlying patterns in SoS, we aim to demonstrate the possibility of automated pattern discovery.

5.1 Environment Construction

We construct a preliminary multi-agent system to simulate a society-level scientific collaboration through an end-to-end pipeline, including collaborator selection, topic discussion, idea generation, novelty assessment, abstract generation, and peer review, inspired by [61, 74, 90]. The overview of our system is shown in Fig. 3. Existing studies primarily focus on simulating individual scientists or small research teams within specific fields (e.g., computer science) and are often constrained to isolated settings that do not capture the broader research ecosystem. In contrast, our work enhances the system's complexity by incorporating realistic factors such as multidisciplinary data (In Appx. C.1), a review and indexing system (In Appx. C.2), and scalable simulation across multiple research teams (In Appx. C.3). More implementation details are provided in Appx. C.4.

5.2 Experiments

Involved Metrics Following the settings of [5, 51, 97], we measure the impact of scientific output by the number of citations a paper receives. In the simulation, the citation counts are updated each time a paper is retrieved during the idea generation phase. For validation, we analyze the citation counts

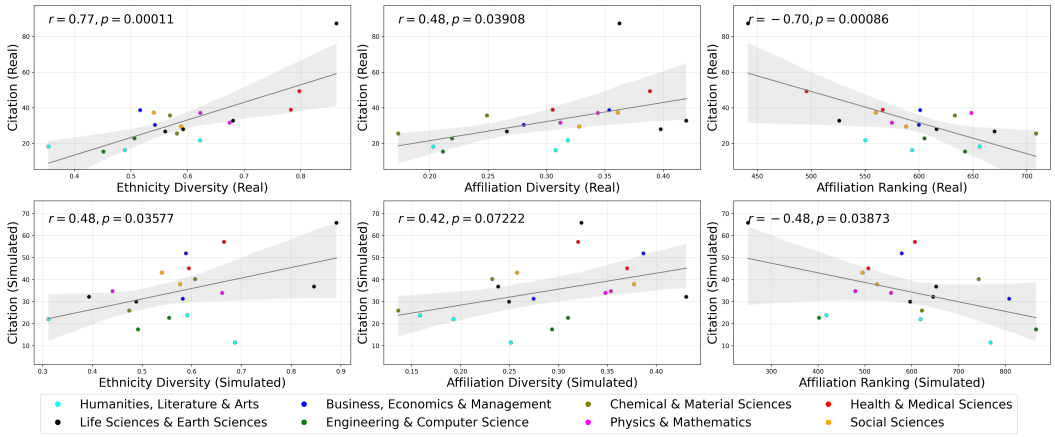

Figure 4: Comparison of real-world (2010) and AI-simulated scientific research patterns. The scatter plots illustrate the relationships between Ethnicity Diversity, Affiliation Diversity, and Affiliation Ranking with Citation Count in both real-world (top row) and simulated (bottom row) data. Correlations observed in real data are partially reproduced by the AI-driven multi-agent system, showing its potential to uncover meaningful research patterns and support automated SoS studies.

of agent-generated papers to assess whether the system can replicate patterns observed in real-world data from the years 2010 to 2011. To evaluate AI's potential in pattern discovery, we examine the influence of three key factors on citation counts: ethnicity diversity, affiliation diversity, and average university ranking. Specifically, we measure diversity using Shannon entropy. For instance, the ethnicity diversity d_{eth} of paper s is calculated as: $d_{eth} = -\sum_{i=1}^{k} p_i(s) \ln p_i(s)$, where k represents the total number of ethnicity categories, and $p_i(s)$ is the proportion of authors from the i-th ethnicity category in paper s.

Simulation Results The experimental results presented in Fig. 4 compare real-world data in 2010 with the outcomes generated by our preliminary LLM-based multi-agent system. Both the real-world and simulated data show that higher citation counts are positively correlated with greater ethnicity diversity, which aligns with existing findings in SoS literature [5], although the correlations are slightly weaker in the simulation. Additionally, the negative correlation between affiliation ranking and citation counts is also reproduced in the simulated data, suggesting that institutions with higher rankings may achieve higher citation counts per research output (a similar comparison using real-world data from 2011 and the simulated results is provided in the Appx. C.5).

Discussions However, while both real-world and simulated data indicate a positive correlation between citation counts and affiliation diversity, the pattern observed in the simulation is not statistically significant (p>0.05). These results suggest that the preliminary AI-driven simulations have the potential to replicate and uncover key patterns in scientific research, but there remains significant room for improvement. For instance, the current system lacks several critical components, such as comprehensive modeling of individual research trajectories and realistic funding and policy influences. These limitations contribute to the preliminary nature of our approach, as the absence of such factors restricts the system's ability to fully capture the complexity of real-world scientific ecosystems. Developing a more comprehensive and sophisticated simulation framework will enhance the system's capability to automatically model complex scientific dynamics with greater accuracy and reliability. More details of outlook are provided in Appx. E.

6 Conclusion

This paper presents a forward-looking perspective on the future of AI4SoS, proposing a five-level autonomy framework toward automated SoS discovery. We show its potential in two critical domains: forecasting trends in technology and innovation, and analyzing the evolution of research communities. We discuss key challenges and future directions, supporting our vision with literature reviews and proof-of-concept studies that showcase early applications. Ultimately, AI4SoS holds the promise of automated SoS discovery, thereby enhancing scientific efficiency and interdisciplinary innovation.

References

[1] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. Nature, pages 1–3, 2024.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[3] Chris Aldrich and Lidia Auret. Unsupervised process monitoring and fault diagnosis with machine learning methods, volume 16. Springer, 2013.

[4] Ahmed Alsayat. Improving sentiment analysis for social media applications using an ensemble deep learning language model. Arabian Journal for Science and Engineering, 47(2):2499–2511, 2022.

[5] Bedoor K AlShebli, Talal Rahwan, and Wei Lee Woon. The preeminence of ethnic diversity in scientific collaboration. Nature communications, 9(1):5163, 2018.

[6] Anurag Ambekar, Charles Ward, Jahangir Mohammed, Swapna Male, and Steven Skiena. Name-ethnicity classification from open sources. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining, pages 49–58, 2009.

[7] Li An, Volker Grimm, Abigail Sullivan, BL Turner Ii, Nicolas Malleson, Alison Heppenstall, Christian Vincenot, Derek Robinson, Xinyue Ye, Jianguo Liu, et al. Challenges, tasks, and opportunities in modeling agent-based complex systems. Ecological Modelling, 457:109685, 2021.

[8] S Balasubramaniam, Vanajaroselin Chirchi, Seifedine Kadry, Moorthy Agoramoorthy, Senthilvel P Gururama, Kumar K Satheesh, and TA Sivakumar. The road ahead: Emerging trends, unresolved issues, and concluding remarks in generative ai—a comprehensive review. International Journal of Intelligent Systems, 2024, 2024.

[9] Luís MA Bettencourt, David I Kaiser, and Jasleen Kaur. Scientific discovery and topological transitions in collaboration networks. Journal of Informetrics, 3(3):210–221, 2009.

[10] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In Conference on fairness, accountability and transparency, pages 149–159. PMLR, 2018.

[11] Taylor Bolt, Jason S Nomi, Danilo Bzdok, and Lucina Q Uddin. Educating the future generation of researchers: A cross-disciplinary survey of trends in analysis methods. PLoS biology, 19(7):e3001313, 2021.

[12] Katy Börner, William B Rouse, Paul Trunfio, and H Eugene Stanley. Forecasting innovations in science, technology, and education. Proceedings of the National Academy of Sciences, 115(50):12573–12581, 2018.

[13] Jinho Chang and Jong Chul Ye. Bidirectional generation of structure and properties through a single molecular foundation model. Nature Communications, 15(1):2323, 2024.

[14] Qian Chen, Yi-Jen Ian Ho, Pin Sun, and Dashun Wang. The philosopher's stone for science–the catalyst change of ai for scientific creativity. Pin and Wang, Dashun, The Philosopher's Stone for Science–The Catalyst Change of AI for Scientific Creativity (March 5, 2024), 2024.

[15] Aaron Clauset, Samuel Arbesman, and Daniel B Larremore. Systematic inequality and hierarchy in faculty hiring networks. Science advances, 1(1):e1400005, 2015.

[16] Jackson De Carvalho and Felix O Chima. Applications of structural equation modeling in social sciences research. American International Journal of Contemporary Research, 4(1):6–11, 2014.

[17] Zhuoyun Du, Chen Qian, Wei Liu, Zihao Xie, Yifei Wang, Yufan Dang, Weize Chen, and Cheng Yang. Multi-agent software development through cross-team collaboration. arXiv preprint arXiv:2406.08979, 2024.

[18] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[19] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[20] Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. ACM Computing Surveys, 55(9):1–33, 2023.

[21] Ashkan Ebadi and Andrea Schiffauerova. How to receive more funding for your research? get connected to the right people! PloS one, 10(7):e0133061, 2015.

[22] Alberto Fernández, Sara del Río, Nitesh V Chawla, and Francisco Herrera. An insight into imbalanced big data classification: outcomes and challenges. Complex & Intelligent Systems, 3:105–120, 2017.

[23] Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S Kohane, and Mihaela van der Schaar. Causal machine learning for predicting treatment outcomes. Nature Medicine, 30(4):958–968, 2024.

[24] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. Science of science. Science, 359(6379):eaao0185, 2018.

[25] Bowen Gao, Bo Qiang, Haichuan Tan, Yinjun Jia, Minsi Ren, Minsi Lu, Jingjing Liu, Wei-Ying Ma, and Yanyan Lan. Drugclip: Contrasive protein-molecule representation learning for virtual screening. Advances in Neural Information Processing Systems, 36, 2024.

[26] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. Large language models empowered agent-based modeling and simulation: A survey and perspectives. Humanities and Social Sciences Communications, 11(1):1–24, 2024.

[27] Alireza Ghafarollahi and Markus J Buehler. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning. arXiv preprint arXiv:2409.05556, 2024.

[28] Charles J Gomez, Andrew C Herman, and Paolo Parigi. Leading countries in global science increasingly receive more citations than other countries doing similar research. Nature Human Behaviour, 6(7):919–929, 2022.

[29] Juan M Górriz, Javier Ramírez, Andres Ortiz, Francisco J Martinez-Murcia, Fermin Segovia, John Suckling, Matthew Leming, Yu-Dong Zhang, Jose Ramón Álvarez-Sánchez, Guido Bologna, et al. Artificial intelligence within the interplay between natural and artificial computation: Advances in data science, trends and applications. Neurocomputing, 410:237–270, 2020.

[30] Roger Guimera, Brian Uzzi, Jarrett Spiro, and Luis A Nunes Amaral. Team assembly mechanisms determine collaboration network structure and team performance. Science, 308(5722):697–702, 2005.

[31] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680, 2024.

[32] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. Interpreting black-box models: a review on explainable artificial intelligence. Cognitive Computation, 16(1):45–74, 2024.

[33] Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. The diversity–innovation paradox in science. Proceedings of the National Academy of Sciences, 117(17):9284–9291, 2020.

[34] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI conference on human factors in computing systems, pages 1–16, 2019.

[35] Iacopo Iacopini, Staša Milojević, and Vito Latora. Network dynamics of innovation processes. Physical review letters, 120(4):048301, 2018.

[36] Guido W Imbens and Donald B Rubin. Causal inference in statistics, social, and biomedical sciences. Cambridge university press, 2015.

[37] Jianchao Ji, Yutong Chen, Mingyu Jin, Wujiang Xu, Wenyue Hua, and Yongfeng Zhang. Moralbench: Moral evaluation of llms. arXiv preprint arXiv:2406.04428, 2024.

[38] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. Bioinformatics, 37(15):2112–2120, 2021.

[39] Justin M Johnson and Taghi M Khoshgoftaar. Survey on deep learning with class imbalance. Journal of big data, 6(1):1–54, 2019.

[40] Yarden Katz and Ulrich Matter. Metrics of inequality: The concentration of resources in the us biomedical elite. Science as Culture, 29(4):475–502, 2020.

[41] Harsurinder Kaur, Husanbir Singh Pannu, and Avleen Kaur Malhi. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM computing surveys (CSUR), 52(4):1–36, 2019.

[42] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT, volume 1, page 2. Minneapolis, Minnesota, 2019.

[43] Seyedmehdi Khaleghian, Himanshu Neema, Mina Sartipi, Toan Tran, Rishav Sen, and Abhishek Dubey. Calibrating real-world city traffic simulation model using vehicle speed data. In 2023 IEEE International Conference on Smart Computing (SMARTCOMP), pages 303–308. IEEE, 2023.

[44] Gohar F Khan, Marko Sarstedt, Wen-Lung Shiau, Joseph F Hair, Christian M Ringle, and Martin P Fritze. Methodological research on partial least squares structural equation modeling (pls-sem) an analysis based on social network approaches. Internet Research, 29(3):407–429, 2019.

[45] Hyunseung Kim, Haeyeon Choi, Dongju Kang, Won Bo Lee, and Jonggeol Na. Materials discovery with extreme properties via reinforcement learning-guided combinatorial chemistry. Chemical Science, 2024.

[46] Gary King, Richard Nielsen, Carter Coberley, James E Pope, and Aaron Wells. Comparative effectiveness of matching methods for causal inference. Unpublished manuscript, Institute for Quantitative Social Science, Harvard University, Cambridge, MA, 2011.

[47] Richard Klavans and Kevin W Boyack. Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? Journal of the Association for Information Science and Technology, 68(4):984–998, 2017.

[48] Mario Krenn and Anton Zeilinger. Predicting research trends with semantic and neural networks with an application in quantum physics. Proceedings of the National Academy of Sciences, 117(4):1910–1916, 2020.

[49] Joffrey L Leevy, Taghi M Khoshgoftaar, Richard A Bauder, and Naeem Seliya. A survey on addressing high-class imbalance in big data. Journal of Big Data, 5(1):1–30, 2018.

[50] Anja K Leist, Matthias Klee, Jung Hyun Kim, David H Rehkopf, Stéphane PA Bordas, Graciela Muniz-Terrera, and Sara Wade. Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. Science Advances, 8(42):eabk1942, 2022.

[51] Weihua Li, Tomaso Aste, Fabio Caccioli, and Giacomo Livan. Early coauthorship with top scientists predicts success in academic careers. Nature communications, 10(1):5170, 2019.

[52] Zuhao Li, Peiran Song, Guangfeng Li, Yafei Han, Xiaoxiang Ren, Long Bai, and Jiacan Su. Ai energized hydrogel design, optimization and application in biomedicine. Materials Today Bio, page 101014, 2024.

[53] Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. Avalonbench: Evaluating llms playing the game of avalon. In NeurIPS 2023 Foundation Models for Decision Making Workshop, 2023.

[54] Zihang Lin, Yian Yin, Lu Liu, and Dashun Wang. Sciscinet: A large-scale open data lake for the science of science research. Scientific Data, 10(1):315, 2023.

[55] Michael Lissack and Brenden Meagher. Navigating the future of large language models in scientific research: Opportunities, challenges, and ethical considerations. Challenges, and Ethical Considerations (September 02, 2024), 2024.

[56] Fengyuan Liu, Talal Rahwan, and Bedoor AlShebli. Non-white scientists appear on fewer editorial boards, spend more time under review, and receive fewer citations. Proceedings of the National Academy of Sciences, 120(13):e2215324120, 2023.

[57] Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, et al. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. Advances in Neural Information Processing Systems, 36, 2024.

[58] Lu Liu, Benjamin F Jones, Brian Uzzi, and Dashun Wang. Data, measurement and empirical methods in the science of science. Nature human behaviour, 7(7):1046–1058, 2023.

[59] Lu Liu, Yang Wang, Roberta Sinatra, C Lee Giles, Chaoming Song, and Dashun Wang. Hot streaks in artistic, cultural, and scientific careers. Nature, 559(7714):396–399, 2018.

[60] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. Information Fusion, 106:102301, 2024.

[61] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. arXiv preprint arXiv:2408.06292, 2024.

[62] Mitra Madanchian and Hamed Taherdoost. Ai-powered innovations in high-tech research and development: From theory to practice. Computers, Materials & Continua, 81(2), 2024.

[63] Mourad Mars. From word embeddings to pre-trained language models: A state-of-the-art walkthrough. Applied Sciences, 12(17):8805, 2022.

[64] Gwenyth Isobel Meadows, Nicholas Wai Long Lau, Eva Adelina Susanto, Chi Lok Yu, and Aditya Paul. Localvaluebench: A collaboratively built and extensible benchmark for evaluating localized value alignment and ethical safety in large language models. arXiv preprint arXiv:2408.01460, 2024.

[65] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6):1–35, 2021.

[66] Lisa Messeri and MJ Crockett. Artificial intelligence and illusions of understanding in scientific research. Nature, 627(8002):49–58, 2024.

[67] Rodrigo Nogueira and Kyunghyun Cho. Passage re-ranking with bert. arXiv preprint arXiv:1901.04085, 2019.

[68] Emma Norling, Bruce Edmonds, and Ruth Meyer. Informal approaches to developing simulation models. Simulating Social Complexity: A Handbook, pages 61–79, 2017.

[69] Kingsley Ofosu-Ampong. Artificial intelligence research: A review on dominant themes, methods, frameworks and future research directions. Telematics and Informatics Reports, page 100127, 2024.

[70] OpenAI. GPT-4 technical report. CoRR, 2023.

[71] Maya L Petersen and Mark J van der Laan. Causal models and learning from data: integrating causal modeling and statistical estimation. Epidemiology, 25(3):418–426, 2014.

[72] Alina Petukhova, Joao P Matos-Carvalho, and Nuno Fachada. Text clustering with llm embeddings. arXiv preprint arXiv:2403.15112, 2024.

[73] Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. On releasing annotator-level labels and information in datasets. arXiv preprint arXiv:2110.05699, 2021.

[74] Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. Large language models as biomedical hypothesis generators: A comprehensive evaluation. arXiv preprint arXiv:2407.08940, 2024.

[75] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. Chatdev: Communicative agents for software development. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pages 15174–15186, 2024.

[76] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140):1–67, 2020.

[77] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pages 429–435, 2019.

[78] Chandan K Reddy and Parshin Shojaee. Towards scientific discovery with generative ai: Progress, opportunities, and challenges. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 28601–28609, 2025.

[79] Guillermo Armando Ronda-Pupo and Thong Pham. The evolutions of the rich get richer and the fit get richer phenomena in scholarly networks: The case of the strategic management journal. Scientometrics, 116(1):363–383, 2018.

[80] Andrey Rzhetsky, Jacob G Foster, Ian T Foster, and James A Evans. Choosing experiments to accelerate collective discovery. Proceedings of the National Academy of Sciences, 112(47):14569–14574, 2015.

[81] Vittoria Scatiggio. Tackling the issue of bias in artificial intelligence to design ai-driven fair and inclusive service systems. how human biases are breaching into ai algorithms, with severe impacts on individuals and societies, and what designers can do to face this phenomenon and change for the better. 2020.

[82] Jule Schulze, Birgit Müller, Jürgen Groeneveld, and Volker Grimm. Agent-based modelling of social-ecological systems: achievements, challenges, and a way forward. Journal of Artificial Societies and Social Simulation, 20(2), 2017.

[83] Reva Schwartz, Leann Down, Adam Jonas, and Elham Tabassi. A proposal for identifying and managing bias in artificial intelligence. Draft NIST Special Publication, 1270, 2021.

[84] Reva Schwartz, Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. Towards a standard for identifying and managing bias in artificial intelligence, volume 3. US Department of Commerce, National Institute of Standards and Technology, 2022.

[85] Feng Shi and James Evans. Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. Nature Communications, 14:1641, 2023.

[86] Feng Shi, Jacob G Foster, and James A Evans. Weaving the fabric of science: Dynamic network models of science's unfolding structure. Social Networks, 43:73–85, 2015.

[87] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In Proceedings of the 24th international conference on world wide web, pages 243–246, 2015.

[88] Diane H Sonnenwald. Scientific collaboration. Annu. Rev. Inf. Sci. Technol., 41(1):643–681, 2007.

[89] Maciej Staszak, Katarzyna Staszak, Karolina Wieszczycka, Anna Bajek, Krzysztof Roszkowski, and Bartosz Tylkowski. Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship. Wiley Interdisciplinary Reviews: Computational Molecular Science, 12(2):e1568, 2022.

[90] Haoyang Su, Renqi Chen, Shixiang Tang, Xinzhe Zheng, Jingzhe Li, Zhenfei Yin, Wanli Ouyang, and Nanqing Dong. Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation. arXiv preprint arXiv:2410.09403, 2024.

[91] Changwon Suh, Clyde Fare, James A Warren, and Edward O Pyzer-Knapp. Evolving the materials genome: How machine learning is fueling the next generation of materials discovery. Annual Review of Materials Research, 50(1):1–25, 2020.

[92] Nees Jan Van Eck and Ludo Waltman. Citation-based clustering of publications using citnetexplorer and vosviewer. Scientometrics, 111:1053–1070, 2017.

[93] Richard Van Noorden and Jeffrey M Perkel. Ai and science: what 1,600 researchers think. Nature, 621(7980):672–675, 2023.

[94] Roberto Verganti, Luca Vendraminelli, and Marco Iansiti. Innovation and design in the age of artificial intelligence. Journal of product innovation management, 37(3):212–227, 2020.

[95] Izhar Wallach, Michael Dzamba, and Abraham Heifets. Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. arXiv preprint arXiv:1510.02855, 2015.

[96] Dashun Wang and Lu Liu. The science of science. In Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, pages 563–564, 2020.

[97] Yang Wang, Benjamin F Jones, and Dashun Wang. Early-career setback and future career impact. Nature communications, 10(1):4331, 2019.

[98] K Hunter Wapman, Sam Zhang, Aaron Clauset, and Daniel B Larremore. Quantifying hierarchy and dynamics in us faculty hiring and retention. Nature, 610(7930):120–127, 2022.

[99] Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. Journal of chemical information and modeling, 59(9):3692–3702, 2019.

[100] Lingfei Wu, Dashun Wang, and James A Evans. Large teams develop and small teams disrupt science and technology. Nature, 566:378–382, 2019.

[101] Stefan Wuchty, Benjamin F Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. Science, 316(5827):1036–1039, 2007.

[102] Shuai Xiao, Junchi Yan, Changsheng Li, Bo Jin, Xiangfeng Wang, Xiaokang Yang, Stephen M Chu, and Hongyuan Zha. On modeling and predicting individual paper citation count over time. In Ijcai, pages 2676–2682, 2016.

[103] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115, 2024.

[104] Yang Yang, Tanya Y Tian, Teresa K Woodruff, Benjamin F Jones, and Brian Uzzi. Gender-diverse teams produce more novel and higher-impact scientific ideas. Proceedings of the National Academy of Sciences, 119(36):e2200841119, 2022.

[105] Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. Oasis: Open agents social interaction simulations on one million agents. arXiv preprint arXiv:2411.11581, 2024.

[106] Yian Yin, Yang Wang, James A Evans, and Dashun Wang. Quantifying the dynamics of failure across science, startups and security. Nature, 575(7781):190–194, 2019.

[107] Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, et al. Chemllm: A chemical large language model. arXiv preprint arXiv:2402.06852, 2024.

[108] Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Evgeny Kharlamov, Bin Shao, et al. Oag: Linking entities across large-scale heterogeneous knowledge graphs. IEEE Transactions on Knowledge and Data Engineering, 35(9):9225–9239, 2022.

[109] Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for LLM agents: A social psychology view. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, pages 14544–14607, 2024.

[110] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. AI open, 1:57–81, 2020.

Appendix

A Related Work

A.1 AI for Science

In recent years, AI has become increasingly common in science and is expected to become the center of research practice [93]. AI has demonstrated great potential to accelerate experimental design, data analysis, optimization problem solving, and discovery of new theories [29, 94, 62]. Specifically, deep neural networks are used to predict the relationship between molecular structures and biological activity [95, 89], reinforcement learning is used to discover unknown materials with superior properties [91, 45], and agent-based systems are introduced to simulate social science scenarios [53, 17]. In addition, as a subfield of science, AI has undergone some preliminary explorations in the SoS [5, 85, 90], revealing promising results.

A.2 Large Language Models

The role of large language models (LLMs) can be articulated from two perspectives: chat (T5 [76], GPT-4 [70], and LLaMA3.1 [19]) and embedding (BERT [42] and DNABERT [38]) generation. First, the capability of dialogue generation enables LLMs to understand user input in natural language and generate contextually relevant responses in various conversational contexts such as knowledge testing, game play, and software programming [107, 57, 109, 17]. Additionally, embedding generation allows LLMs to convert input text into fixed-dimensional vector representations, which effectively capture the semantic information of the text and can be used for tasks such as text similarity computation, information retrieval, and sentiment analysis [67, 63, 4, 72]. Therefore, the capabilities of LLMs in both text generation and embedding generation make them applications spanning from natural language processing tasks to more complex domains such as SoS, where they can assist in understanding research dynamics, scientific discovery, and scientific collaboration.

B Datasets for Science of Science Research

Currently, there are several large-scale, cross-disciplinary academic datasets for SoS research: Microsoft Academic Graph (MAG), Open Academic Graph (OAG), and SciSciNet, where the statistical information of each dataset is summarized in Table 2.

Table 2: Summary table of large-scale cross-discipline academic datasets.

Datasets	MAG	OAG	SciSciNet
Due	2020	2023	2021
Domain	Art, Biology, Business, Chemistry, Computer Science, Economics, Engineering, Environmental Science, Geography, Geology, History, Materials Science, Mathematics, Philosophy, Physics, Political Science, Psychology, Sociology	Art, Biology, Business, Chemistry, Computer Science, Economics, Engineering, Environmental Science, Geography, Geology, History, Materials Science, Mathematics, Philosophy, Physics, Political Science, Psychology, Sociology	Art, Biology, Business, Chemistry, Computer Science, Economics, Engineering, Environmental Science, Geography, Geology, History, Materials Science, Mathematics, Medicine, Philosophy, Physics, Political Science, Psychology, Sociology
Author	261,445,825	35,774,510	134,197,162
Paper	247,389,875	130,710,733	134,129,188
Affiliation	25,811	143,749	26,998

C More Experimental Details

C.1 Multidisciplinary Data

We use the OAG 3.1[1] as the initial database for our system, which developed from the Open Academic Graph [108]. This data set includes 35,774,510 authors and 130,710,733 papers as of 2023, spanning diverse domains such as physics, chemistry, and computer science. In Table 3, we present the disciplines and fields of paper in the Open Academic Graph, which is used to analyze the potential different patterns in various areas. We use papers from 2002 to 2009 as the reference database and papers from 2010 to 2011 as the validation database. To address missing author ethnicity and paper field information—key elements for validating SoS findings—we employ several data completion strategies. Specifically, we adopt corresponding approaches for the various pieces of author information and paper information in this dataset for our simulation, shown in Table 4 and 5.

Table 3: Summary table of disciplines and fields [5].

Field	Discipline
Humanities, Literature & Arts	[Art, History, Philosophy, Psychology]
Life Science & Earth Sciences	[Biology, Environmental Science, Geography, Geology]
Business, Economics & Management	[Business, Economics]
Engineering & Computer Science	[Computer Science, Engineering]
Chemical & Material Sciences	[Chemistry, Materials Science]
Physics & Mathematics	[Mathematics, Physics]
Health & Medical Sciences	[Medicine]
Social Sciences	[Political Science, Sociology]

[1] https://open.aminer.cn/open/article?id=65bf053091c938e5025a31e2

Table 4: Different strategies are adopted for various pieces of information regarding authors.

Field Name	Strategy	Example
	Author Information	
Name	Use the anonymization technique	Scientist 1
Ethnicity	Use the name ethnicity classifier [6]	British
Affiliation	Retain the original content	[King's College London]
Affiliation Ranking	Use THE World University Rankings 2025 [1]	36
Citation	Extract the author's published papers between 2010 to 2020 and calculate the total number of citations for the papers; In the simulation, it will be updated if his/her paper is cited	1800
Co-author	Extract the author's published papers between 2010 to 2020 and record the collaborators in the papers; In the simulation, it will be updated if there are new collaborators	[Scientist 10, Scientist 201, Scientist 1002, . . .]
Discipline	Extract the author's published papers between 2010 to 2020 and assign the author's discipline as the one that appears most frequently	Psychology
Research topic	Extract the author's published papers between 2010 to 2020 and record the keywords in the papers; Use GPT-4 to summarize these keywords into research topics	[Neuropsychology, Cognitive flexibility, Attentional bias, . . .]

[1] `https://www.timeshighereducation.com/world-university-rankings/latest/world-ranking`

C.2 Review and Indexing System

To better simulate and reveal the patterns of scientific collaboration mechanisms, we introduce a review and indexing system. Papers written by scientist teams are peer-reviewed and scored (ranging from 1 to 10), and those that exceed the acceptance threshold (with score larger than 5) are added to the reference paper database as newly published papers. In Table 6 and 7, we present the peer review criteria used in our simulation system, which is based on the modified Neural Information Processing Systems review guidelines [2] considering that the papers produced by cross-discipline agents are not all in the field of computer science. Although this criteria comes from a computer science conference, the basic evaluation metrics can be applied in multiple areas. Besides, the indexing system allows agents to retrieve published papers as references, and the citation count of referenced papers is updated accordingly, which is later used for metric evaluation.

C.3 Scalable Simulation

To better replicate the phenomenon of free collaboration in real scientific cooperation, we implement an adaptive concurrent distributed system based on the OASIS [105]. The system's asynchronous mechanism achieves concurrent processing by queuing multiple requests from agents in an inference channel and then distributing them to different ports for sending and receiving, where each port has deployed an LLM responsible for chatting or embedding. Furthermore, to reduce CPU load, we set the channel allocation wait time based on the number of pending requests in the channel, thereby enabling long-term large-scale asynchronous simulation. This mechanism serves the two purposes: 1. Enabling scientist agents from different teams to communicate simultaneously, including both intra-team and cross-team collaboration, and 2. Accelerating the simulation process to enable large-scale simulations at the million-agent level. We test the time cost of our simulation system under different number of agents, illustrated in Fig. 5. It could be found that we realize a fast large-scale agent system, where a simulation of a million agent society takes only one week.

[2] `https://neurips.cc/Conferences/2024/ReviewerGuidelines`

Table 5: Different strategies are adopted for various pieces of information regarding papers.

Field Name	Strategy	Example
	Paper Information	
Title	Retain the original content	Linkages of plant traits to soil properties ...
Abstract	Retain the original content	Global change is likely to alter plant community ...
Year	The year of the papers in the initial database is set to -1, while the papers published by the agent are assigned the epoch when the review is accepted	-1
Citation	In the initial database, the citation count of the papers is the original citation value plus the number of times they are cited during the simulation, while the citation count of the papers written by the agent is the number of times they are cited during the simulation	82
Authors	Retain the original content	[Scientist 124, Scientist 7923, ...]
Cited Paper	The papers in the initial database have None for this information due to its absence, while the papers published by the agent contain the names of the cited papers	None
Discipline	Use GPT-4 to classify the papers into disciplines based on their keywords and titles. Refer to Table 3 for all the disciplines used	Environmental Science

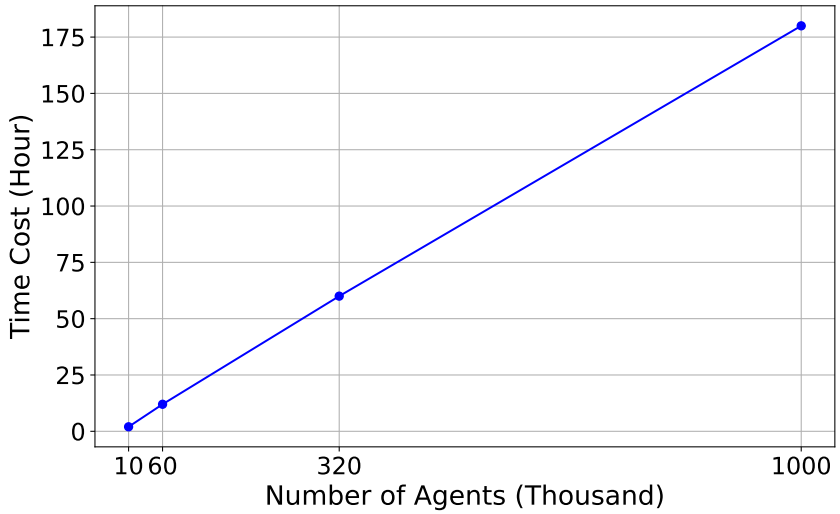

Figure 5: The time taken for a complete scientific collaboration with agents of different scales. A simulation of a million-agent society takes only one week.

C.4 Implementation Details

We implement our system on 32 NVIDIA A100 GPUs, with 4 ports deployed on each GPU, and each port running the *LLaMA3.1-8b* model. We allow each agent to create up to 3 teams simultaneously, with team sizes following an exponential distribution. This is because we analyze the team sizes of papers published between 2002 and 2009 in the OAG (over 1,000,000 papers), as shown in Fig. 6. The red fitting line indicates that the team sizes in the real data follow an exponential distribution. Therefore, in our simulation, the team size of each agent is also modeled using an exponential distribution.

Table 6: Prompt Tailored for Multidisciplinary Reviewers

Prompt Tailored for Multidisciplinary Reviewers (1/2)
You are a researcher from a multidisciplinary background reviewing a paper that has been submitted to a venue that involves multiple scientific disciplines. Be critical and cautious in your decision-making. If the paper has significant weaknesses or you are uncertain about its quality, provide lower scores and recommend rejection. Below are the questions you will be asked on the review form for each paper and some guidelines on what to consider when answering these questions. Reviewer Guidelines for Multidisciplinary Paper Review: 1. Summary: Provide a brief summary of the paper and its contributions. This is not the place to critique the paper. The authors should generally agree with a well-written summary, which reflects an accurate understanding of their work from a multidisciplinary perspective. 2. Strengths and Weaknesses: Please provide a thorough assessment of the strengths and weaknesses of the paper, touching on each of the following dimensions: - Originality: Are the tasks or methods novel within each of the relevant disciplines? Does the work represent an innovative combination of techniques or concepts from different fields? Is it clear how this work distinguishes itself from previous contributions in each discipline involved? - Quality: Is the submission technically sound in each of the relevant fields? Are claims well-supported by evidence (e.g., theoretical analysis or experimental results)? Are the methods used appropriately for each discipline involved? Is this a complete piece of work, or still a work in progress? Are the authors transparent and honest in evaluating both the strengths and weaknesses of their work? - Clarity: Is the paper written in a way that is accessible to readers from multiple disciplines? Is it well-organized, with clear explanations of concepts across different fields? If not, please suggest improvements for clarity. Does it provide sufficient detail for an expert in each relevant field to understand the methodology and reproduce results? - Significance: Are the results important? Are others (researchers or practitioners) likely to use the ideas or build on them? Does the submission address a difficult task in a better way than previous work? Does it advance the state of the art in a demonstrable way? Does it provide unique data, unique conclusions about existing data, or a unique theoretical or experimental approach? 3. Questions: Please list any questions or suggestions that could help clarify the paper's limitations or improve its quality. Responses from the authors could change your opinion or address areas of confusion. This feedback can be critical for the rebuttal and discussion phase with the authors.

In idea generation and novelty assessment, each agent can cite up to 9 references per speech, where the retrieval results are obtained based on the similarity between the embeddings of the query terms and the embeddings of the papers in the database. The model used for embedding is *mxbai-embed-large*. To avoid storage issues, each agent's memory retains a maximum of 5 entries. Each paper undergoes peer review by 3 reviewers. In terms of the timeline, each epoch allows for 1 action, meaning a complete scientific collaboration can be completed in 6 epochs if the team progresses without any delays or interruptions. In our final experiment, the size of our society is maintained at 1 million agents, with a total of 40 epochs.

C.5 More Experimental Results

A similar comparison using real-world data from 2011 and the simulated result is provided in Fig. 7. The statistical analysis of the 2011 data exhibits similar trends to those observed in Fig. 4, which presents the comparison using 2010 data. The positive correlation between citation counts and ethnicity diversity, as well as the negative correlation between affiliation ranking and citation counts, are consistently reflected in both years. However, minor variations in correlation strength are observed, highlighting the dynamic nature of scientific collaboration trends over time.

D Alternative Views

The application of AI in SoS is often seen as transformative, promising to accelerate discovery. However, critics highlight significant limitations and risks, questioning its unqualified benefits. These concerns focus on systemic issues and unintended consequences [10, 77, 66]. Key counterarguments

Table 7: Prompt Tailored for Multidisciplinary Reviewers

Prompt Tailored for Multidisciplinary Reviewers (2/2)
4. Ethical Concerns: Flag any ethical concerns, particularly those that may arise from interdisciplinary collaboration. Ensure any ethical issues related to research design, data usage, or broader implications are addressed. 5. Overall Score: Provide a final score based on the paper's strengths and weaknesses. Use the following scale: - 10: Award Quality: A technically flawless paper with groundbreaking impact across one or more disciplines, with exceptionally strong evaluation, reproducibility, and resources, and no unaddressed ethical concerns. - 9: Very Strong Accept: A technically flawless paper with groundbreaking impact in at least one area and strong impact on multiple areas, with flawless evaluation, resources, and reproducibility, and no unaddressed ethical concerns. - 8: Strong Accept: A technically strong paper with novel ideas, significant impact on at least one discipline or moderate-to-high impact on multiple areas, with excellent evaluation, resources, and reproducibility, and no unaddressed ethical concerns. - 7: Accept: A technically solid paper with moderate-to-high impact in one or more subfields, good-to-excellent evaluation, reproducibility, and resources, and no unaddressed ethical concerns. - 6: Weak Accept: A solid paper with moderate impact, no major concerns in terms of evaluation, reproducibility, and ethical considerations. - 5: Borderline Accept: A technically solid paper where reasons to accept outweigh reasons to reject, e.g., limited evaluation. Use sparingly. - 4: Borderline Reject: A technically solid paper where reasons to reject outweigh reasons to accept, e.g., limited evaluation. Use sparingly. - 3: Reject: A paper with technical flaws, weak evaluation, inadequate reproducibility, or incompletely addressed ethical concerns. - 2: Strong Reject: A paper with major technical flaws, poor evaluation, limited impact, poor reproducibility, or mostly unaddressed ethical considerations. - 1: Very Strong Reject: A paper with trivial results, poor evaluation, or unaddressed ethical issues.

include: (1) Reinforcement of Existing Inequalities: AI systems rely heavily on historical data, which often mirror long-standing inequities within the scientific community. For instance, datasets may disproportionately represent well-established disciplines, regions, or researchers, thereby perpetuating an imbalanced view of scientific contributions. Critics argue that this could stifle innovation by overlooking emerging fields and underrepresented groups, ultimately reinforcing the leading trend rather than fostering diversity. (2) Overreliance on Traditional Metrics: Academic evaluation metrics, such as citation counts and journal impact factors, are central to many AI applications in SoS. These metrics have been criticized for prioritizing mainstream research while marginalizing unconventional or nascent ideas. Opponents caution that AI-driven analyses might amplify this bias, narrowing the scope of scientific discovery and undervaluing novel contributions.

While these critiques highlight significant challenges, they underscore the importance of addressing fairness, and inclusivity in AI applications for SoS [34, 83, 84]. To mitigate these concerns, the following strategies can be adopted: (1) Promoting Diversity in Data and Metrics: Expanding data curation efforts to include a wider range of disciplines, regions, and research communities is critical for minimizing biases. Additionally, developing diversified scientific impact metrics beyond citation counts can ensure a more equitable evaluation of research contributions. (2) Incorporating Bias Mitigation Techniques: Embedding bias detection and correction mechanisms in AI systems can help identify and address inequities in the data and algorithms. These techniques should be complemented by rigorous validation to ensure fairness and reliability.

E Outlook

As AI4SoS progresses toward full autonomy, we envision a future where scientific discovery itself becomes a more self-reflective, adaptive, and strategically guided process. In this envisioned landscape, AI agents are trained on vast corpora of scholarly data and historical innovation patterns, which will

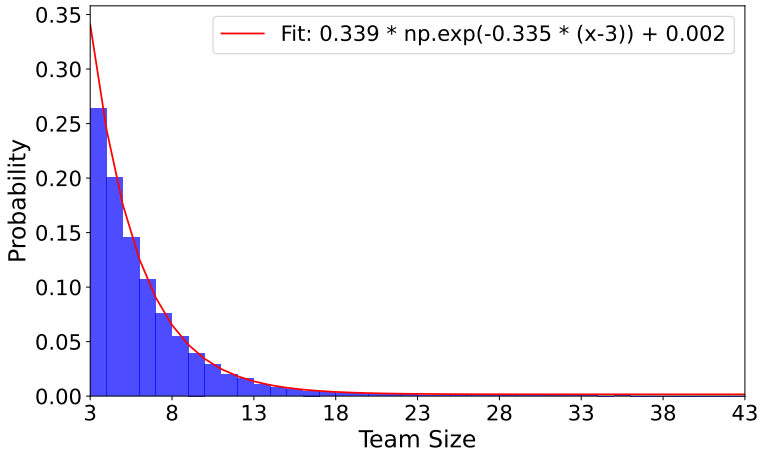

Figure 6: The statistics of team sizes for papers published between 2002 and 2009 in the OAG, with the red fitting line revealing that the distribution follows an exponential pattern.

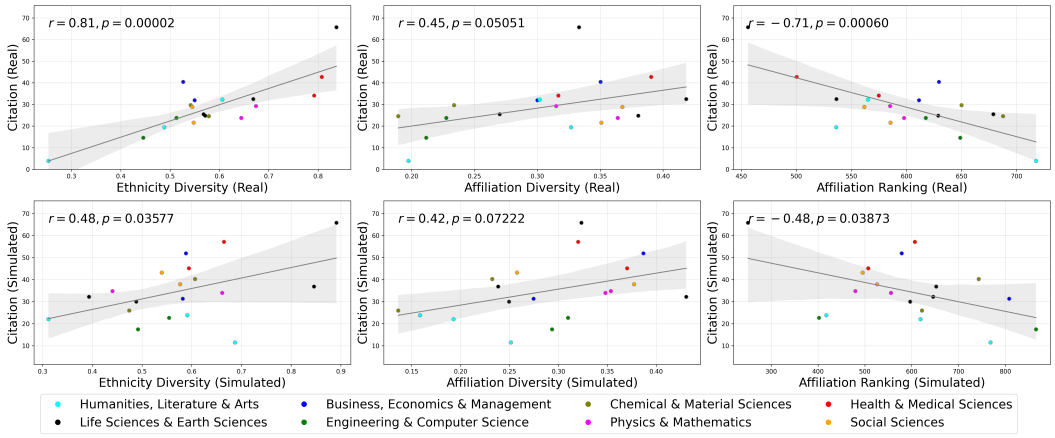

Figure 7: Comparison of real-world (2011) and AI-simulated scientific research patterns.

not only map the contours of scientific fields but also anticipate emerging disciplines and recommend actionable research agendas.

Automated SoS systems will continuously monitor the evolving structure of scientific collaboration, offering dynamic guidance to policymakers, institutions, and individual researchers. Research teams may be formed or optimized based on predicted synergy and complementary expertise, while funding strategies could adapt in real time to maximize long-term innovation impact. Moreover, AI4SoS could democratize scientific foresight, making sophisticated analyses accessible to a broader range of stakeholders, from early-career researchers to global research organizations. The resulting ecosystem would be one where science is not only accelerated but also made more transparent, inclusive, and responsive to societal needs.

To enhance real-world applicability, we also envision deployment scenarios in which AI4SoS integrates directly with existing scientific ecosystems. For instance, it could serve as a sandbox environment for evaluating national research policies, allowing simulated assessments before implementation. Within academic institutions, AI4SoS could support internal research strategy formulation, identifying growth areas and optimizing resource allocation. Additionally, it could assist governmental and funding bodies in planning emerging discipline layouts and national innovation agendas. These integration pathways would significantly boost the practical value, societal impact, and credibility of AI4SoS.

Achieving this vision will demand sustained interdisciplinary collaboration, ethical oversight, and robust infrastructure, but the potential payoff is immense: a future in which the SoS is not just studied, but actively shaped by intelligent systems.

F Impact Statement

We believe that sustained collaboration between AI researchers and SoS scholars is essential for advancing our understanding of complex scientific processes. This study leverages the complementary expertise of both fields to address key SoS challenges, improving scientific efficiency and fostering interdisciplinary innovation.

However, from an ethical perspective, the integration of AI with SoS research may present several concerns. First, **accountability:** When AI participates in scientific decision-making, it is crucial to clarify responsibility. For instance, if an AI-generated prediction leads to errors, should developers bear full responsibility? We suggest enhancing AI system transparency (e.g., recording decision-making pathways) and explainability (e.g., providing reasoning behind decisions) to help researchers and regulators delineate accountability more clearly. Second, **fairness and bias:** AI systems rely on training data, which may contain inherent biases related to gender, geography, or economic disparities. These biases can lead to unjust scientific conclusions. Therefore, AI development and application should include rigorous data preprocessing and incorporate fairness constraints within algorithms to mitigate the risk of bias propagation. Finally, **public trust:** AI-driven automation tools, due to their complexity, may create a sense of detachment among the public. When AI decision-making processes are opaque, concerns about the credibility of scientific findings may arise. To foster trust, it is essential to develop more interpretable AI models and ensure human oversight in scientific processes.

From a societal perspective, the complexity of SoS demands innovative approaches. Conventional statistical studies, which depend largely on historical data, frequently struggle to uncover causal mechanisms. In contrast, agent-based AI provides a dynamic, causality-driven alternative. By elucidating the mechanisms behind the evolution of scientific knowledge, these methods can clarify how government policies influence research funding, academic publishing, and interdisciplinary collaboration. As AI4SoS advances, it will foster more effective knowledge exchange among academia, industry, and government, accelerating technological and theoretical innovation. Through intelligent analysis and predictive modeling, researchers can more precisely identify scientific challenges, significantly enhancing the efficiency of discovery.