A SURVEY OF HUMAN-AI COLLABORATION FOR SCIENTIFIC DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Artificial intelligence (AI) is increasingly integrated into scientific discovery processes, such as protein design, gene analysis, and materials research, significantly improving research efficiency in these fields. While much of the recent literature emphasizes fully automated pipelines, it is crucial to acknowledge that scientific discovery is inherently a creative and high-stakes endeavor. Therefore, it relies heavily on human expertise for judgment and guidance, especially in the face of uncertainty. Despite the rapid growth in human-in-the-loop and collaborative systems, the field still lacks a unifying survey that explains how humans and AI collaborate throughout the scientific discovery life cycle. In this paper, we present a systematic review of human–AI (HAI) collaboration for scientific discovery. Specifically, we identify four representative roles that humans and AI can assume, including *Informer, Explorer, Evaluator*, and *Controller*. Using this lens, we then distill common HAI collaboration patterns across three distinct stages in the scientific discovery process (i.e., *observation*, *hypothesis*, and *experiment*). Finally, we identify key gaps in existing approaches and outline future research directions for developing trustworthy, role-aware HAI collaboration systems in scientific discovery.

1 INTRODUCTION

The integration of artificial intelligence (AI) into scientific research has progressed from a methodological enhancement to a fundamental paradigm shift. Early AI models primarily functioned as computational tools, facilitating low-level analytical tasks such as pattern extraction and representation learning (Wang et al., 2023a), which later specialized into domain-specific scientific models for tasks like protein and molecular modeling (Kwon et al., 2024). Recently, advances in AI, particularly large language models (LLMs), have introduced stronger reasoning capabilities (Ma et al., 2024), exerting a transformative effect on scientific discovery. For example, some systems even demonstrate the ability to plan and execute experiments autonomously (Zheng et al., 2025).

Despite the rapid development of AI, scientific discovery remains a fundamentally creative and complex process that requires significant human involvement. Especially in high-stakes and resource-intensive scientific domains (e.g., medicine, chemistry, and genomics) where errors can be costly or irreversible, human scientists are still expected to continuously monitor AI outcomes and make critical research decisions. However, existing surveys predominantly focus on the technical capabilities of AI models, often overlooking the role of human scientists in the discovery process. For instance, Zheng et al. (2025) review existing LLM-based systems for scientific discovery and categorize them into three levels of AI autonomy (i.e., Tool, Analyst, and Scientist). Reddy & Shojaee (2025) provide a survey on the applications of generative AI across scientific tasks and summarize challenges in developing comprehensive AI systems for scientific discovery. Consequently, the theoretical understanding of how humans and AI can effectively collaborate throughout the scientific research process remains limited. While there are existing surveys focusing on general human-AI (HAI) interaction and discussing relevant interaction principles (Huang et al., 2025; Li et al., 2024), they often overlook the specific context of scientific discovery. Therefore, there is a lack of a structured framework for characterizing the HAI partnerships in scientific discovery.

To bridge this gap, we presented a systematic survey of HAI collaboration in scientific discovery. First, we identified four roles (*Informer, Explorer, Evaluator, Controller*) of human and AI in collaborative

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

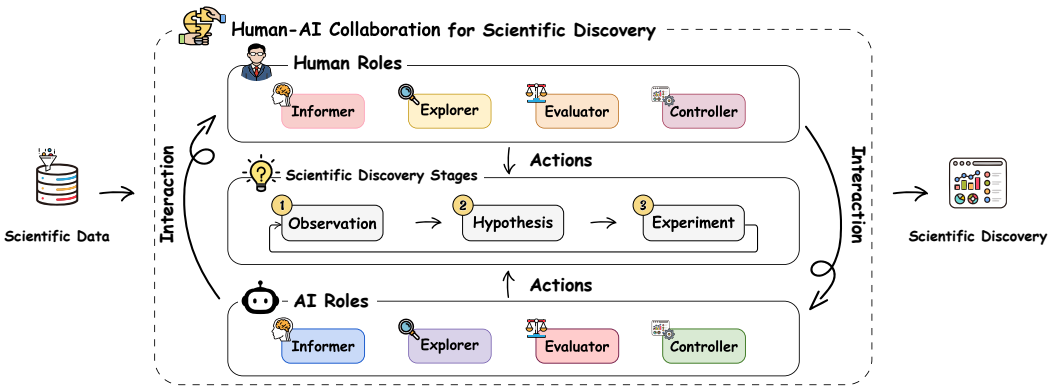


Figure 1: Our taxonomy characterizes research on HAI collaborative scientific discovery from four roles of human and AI across the three stages of scientific discovery.

systems based on a systematic review of 65 papers. We anchored our analysis in the established three stages of scientific discovery (Wang et al., 2023a) (*observation*, *hypothesis*, and *experiment*), as illustrated in Figure 1. Building upon this taxonomic framework, we analyzed the specific role allocation between human and AI partners, identifying and discussing their common collaboration patterns in scientific discovery and their role differences across the three stages. We concluded by outlining five open challenges and future directions for establishing efficient and trustworthy HAI partnerships. The main contributions of this paper are summarized as follows:

- We presented a comprehensive survey that systematically reviews HAI collaboration specifically in scientific discovery.
- We introduced a novel taxonomy that defines four roles of human and AI and characterized their distinct collaboration patterns across the three stages of scientific discovery.
- We identified critical challenges and future pathways for building effective and trustworthy HAI partnerships in the scientific discovery process.

2 Methodology

In this section, we outline the methodology for collecting a corpus of 65 papers on HAI collaboration for scientific discovery and the coding process used to identify the roles of human and AI.

2.1 Paper Collection

To assemble a high-quality corpus, we systematically searched for studies published between January 2016 and June 2026. Since relevant systems span AI, human-computer interaction, visualization, and natural sciences, we searched the main venues of these communities, including NeurIPS, ICML, ICLR, AAAI, IJCAI, ACL, EMNLP, NAACL, CHI, UIST, CSCW, IEEE VIS, IEEE TVCG, EuroVis, Nature, and Science. We conducted a keyword search by pairing terms related to scientific discovery and scientific domains (e.g., "scientific discovery", "scientific research", "medicine", and "chemistry") with HAI collaboration terms (e.g., "human-AI collaboration", "human-in-the-loop", and "AI assistant"), yielding 264 records. We then augmented these records through iterative backward and forward snowballing over the retrieved papers and two authoritative surveys on AI for science (Zheng et al., 2025; Reddy & Shojaee, 2025). This process continued until no further relevant studies emerged, adding 32 records in total. After removing duplicates, three co-authors independently read the remaining 209 records and resolved all disagreements through discussion. In the end, we retained 65 papers (details in Appendix C) that satisfied all of our inclusion criteria.

We included a paper only if it met three criteria. 1) It presents an interactive system or workflow explicitly designed to facilitate scientific discovery. We excluded fully automated pipelines because our focus is the division of labor between humans and AI rather than autonomous algorithms. 2) Its contribution extends beyond data labeling and annotation. We excluded works whose only output is a

dataset, such as SciDaSynth (Wang et al., 2025b), which interactively extracts structured evidence tables from scientific literature, even when the authors argued that it could support later scientific discovery, because such tools operate before rather than within the discovery process. 3) It met a quality threshold, being either peer-reviewed or, if a preprint, cited more than 100 times, thereby filtering out immature or unvetted work. Appendix A details the paper collection process.

2.2 Paper Coding

To analyze how humans and AI collaborate in the scientific discovery process, six co-authors conducted an inductive coding process to derive the roles of humans and AI from the corpus. First, each coder independently coded a subset of papers and labeled every human or AI actor with a short description of its function in the collaboration. We then grouped these descriptions by functional similarity, merged overlapping categories, and refined their boundaries through weekly discussions. After several rounds of comparison, the candidate labels converged on four recurring functions: organizing existing evidence, generating or searching for new research directions or artifacts, judging candidate validity, and steering the collaborative process. We named these functions *Informer, Explorer, Evaluator*, and *Controller*, which covered every actively collaborating human or AI role in the corpus. During coding, we also examined alternative ways of grouping these functions. For example, we considered merging evidence organization and candidate generation into a single role. However, such a scheme failed to capture an important difference in collaborative agency: some actors make existing information interpretable and actionable for others, whereas others actively expand the research space by proposing new patterns, hypotheses, or designs.

After finalizing the role codebook, we applied it to the full corpus of 65 papers. To examine how these roles operate across the scientific discovery process, we also coded each paper using an established three-stage decomposition of scientific discovery (Wang et al., 2023a): *observation, hypothesis*, and *experiment*. Specifically, the *observation* stage involves collecting and examining data or phenomena to identify patterns, anomalies, or open questions that require explanation. Based on these observations and prior knowledge, researchers formulate research hypotheses in the *hypothesis* stage, which are tentative and testable explanations or predictions that guide subsequent inquiry. The *experiment* stage then involves designing and conducting studies or analyses to test these hypotheses, using the results to validate, refine, or reject them, often leading to new observations and continuing the scientific discovery cycle.

Each paper was assigned to three of the six co-authors, who coded it independently using the shared codebook. We measured inter-coder agreement with Krippendorff's α, obtaining 0.78 for role coding and 0.85 for stage coding, suggesting reliable agreement. When disagreements occurred, each coder first reviewed the conflicts and adjusted their codes to address apparent issues, such as missing or inconsistent codes. Unresolved disagreements were then discussed in weekly meetings until agreement was reached. We recorded all resolutions in the codebook to maintain consistency in subsequent coding. Details of the codebook are in Appendix B.

3 Taxonomy

In this section, we first introduce four roles that humans or AI can play in the scientific discovery process. Built upon the definitions of these roles, we then elaborate on common HAI collaboration patterns at the three stages of scientific discovery. Finally, we analyze how these roles differ across the three distinct stages.

3.1 Roles of Human and AI

Based on our systematic analysis of the corpus, we identify four roles of human and AI in scientific discovery: **Informer**, **Explorer**, **Evaluator**, and **Controller**. To clarify agency, we apply "human-" or "AI-" prefixes before these roles (e.g., human-Informer, AI-Informer).

$\diamond$ **Informer**. The *Informer* synthesizes, distills, or articulates key information, insights, or constraints from raw data, intermediate analyses, or simulated processes to guide the actions of other roles. For instance, the AI-Informer in THALIS extracts temporal patterns from longitudinal symptom records in cancer therapy (Floricel et al., 2022). This provides a summarized trajectory view for experts to

assess patient responses and decide on the subsequent treatment plan. Similarly, in ISHMAP for Mars rover operations (Wright et al., 2023), the human-Informer marks instrument states and operational events on the telemetry timeline. The AI uses these annotations to reduce false alarms during state changes and to highlight unexpected signals.

◇ **Explorer**. The *Explorer* operates within the space of data patterns, hypotheses or experimental designs to explore promising candidates or directions. In contrast to the Informer, the *Explorer* searches for or generates new candidate artifacts tailored to each stage. Thus, a system would be coded as an AI-Informer when it summarizes existing patient records, but as an AI-Explorer when it further proposes candidate treatment pathways for human validation. For instance, in the hAE interface (Pratiush et al., 2024), the AI-Explorer searches the parameter space to select the next experimental conditions for electron microscopy. ChemVA (Sabando et al., 2021) enables the human-Explorer to interactively navigate a projected chemical space and select promising candidate molecules to be validated in subsequent experiments.

◇ **Evaluator**. Once the data patterns, hypotheses or experimental designs are proposed, their scientific merit must be rigorously evaluated and even refined. The *Evaluator* assesses these artifacts based on predefined criteria, evidence, and domain constraints, deciding whether they are worth keeping and revising them when necessary to meet quality standards. For instance, in RetroLens (Shi et al., 2023a), chemists serve as the human-Evaluator. Specifically, they can assess AI-predicted synthetic routes for chemical feasibility and, if necessary, refine the synthetic steps by themselves.

◇ **Controller**. The *Controller* oversees the whole scientific discovery workflow to ensure correct and constraint-compliant execution, intervening when necessary to adjust procedures and handle runtime exceptions. Unlike the Evaluator, the *Controller* governs the process rather than judging the scientific merit of an artifact; it decides what runs next, sequences steps, adjusts parameters, or recovers from failures. For example, a chemist is coded as a human-Evaluator when judging the feasibility of a synthetic route (Shi et al., 2023a), but as a human-Controller when deciding how the molecular synthesis workflow should proceed (Darvish et al., 2025). Additionally, in ORGANA (Darvish et al., 2025), the AI-Controller operates robotic laboratory equipment and executes experimental procedures, while adapting low-level actions to support the successful completion of chemistry experiments.

3.2 Common Collaboration Patterns Within Each Stage of Scientific Discovery

In this section, we introduce common HAI collaboration patterns at each of the three stages (i.e., *observation, hypothesis*, and *experiment*) of the scientific discovery process. Note that although many papers in our corpus involve multiple human or AI roles, we only focus on roles that actively participate in HAI collaboration and derive collaboration patterns from them to ensure that our analysis reflects meaningful HAI collaboration rather than mere role co-existence.

Observation Stage

The *observation* stage involves collecting and analyzing data to identify patterns and anomalies that warrant further investigation. During this stage, humans and AI collaborate to organize large datasets, highlight potential patterns, and verify them against the raw data to support hypothesis generation.

AI-Informer & Human-Explorer. A common HAI collaboration pattern in the *observation* stage involves an AI-Informer transforming raw data into structured representations (e.g., embeddings, feature importance maps), enabling a human-Explorer to identify clusters, trends, and outliers in scientific data. For example, in biological domains, AI-Informers can visualize tissue interactions (Mörth et al., 2025) or cell trajectories (Wang et al., 2026), and human-Explorers follow these visual pathways to uncover disease trends or developmental stages. AI-Informers can also organize data by similarity or explanatory cue. For instance, they cluster compounds or highlight sequence motifs, allowing human-Explorers to search for promising chemicals (Sabando et al., 2021) or protein functions (Park et al., 2024). Also, they can rank the influence of features on predictions, with human-Explorers investigating these cues to understand air pollution drivers (Palaniyappan Velumani et al., 2022), explore raw fiber tracts (Xu et al., 2023), and analyze phenotype images (Krueger et al., 2020). Recent work extends the AI-Informer role from organizing observed data to producing simulated observations for human exploration. In social media event analysis, the AI-Informer role-plays social media users and generates their evolving responses, thereby creating simulated event

trajectories for analysis. The human-Explorer then examines the trajectories to trace how the events unfold and understand what factors drive their evolution (Cheng et al., 2025).

AI-Explorer & Human-Evaluator. When models can automatically suggest candidates, the collaboration often follows an examination by humans. The AI-Explorer suggests candidate patterns and displays the primary evidence it used, such as highlighted inputs, matched records, or similar past cases. The human-Evaluator reviews these candidates to determine whether they are correct and meaningful. For instance, in drug research, the AI-Explorer searches a biomedical knowledge graph and proposes drug-disease links that were not previously established, while the human-Evaluator traces the supporting paths to decide whether each link is biologically credible enough to keep (Wang et al., 2023b). For longitudinal records in medicine, the AI-Explorers find distinct treatment pathways or care rules. The human-Evaluators review these specific sequences to validate symptom progression or hospital protocols (Floricel et al., 2022). In plant embryo lineage analysis, the AI-Explorer can generate classification results from multiple models. The human-Evaluator then assesses these outputs to identify the correct cell type based on consensus (Hong et al., 2024). Additionally, the AI-Explorer proposes candidate functions for a gene set and checks each claim against curated databases itself, while the human-Evaluator judges which functions are relevant and well supported (Wang et al., 2025c).

AI-Explorer & Human-Controller. When observation requires iterative search over a large data or evidence space, the collaboration often pairs a human-Controller with an AI-Explorer. The human-Controller directs the search process, while the AI-Explorer carries out the search and returns candidate patterns or evidence for further inspection. For literature discovery, the human-Controllers adjust search constraints and decide which direction the search should follow, while the AI-Explorers navigate databases and return a ranked set of candidate papers for subsequent screening (Qiu et al., 2025). To identify star clusters or spatial groups, the human-Controller adjusts the detection parameters. The AI-Explorer accordingly scans astronomical surveys to locate new star clusters (Ratzenböck et al., 2023) or analyzes spatial population data to find notable spatial groups (Wentzel et al., 2023). Additionally, for Mars rover operations, the human-Controller sets anomaly criteria to enable the AI-Explorer to surface candidate anomaly events in rover telemetry, which become new operational signals for mission experts to investigate (Wright et al., 2023).

HYPOTHESIS STAGE

Hypothesis generation involves proposing explanations or solutions for observed phenomena. Collaborative efforts during this phase focus on retrieving background knowledge and developing candidate theories or designs to guide subsequent testing.

AI-Explorer & Human-Evaluator. A common collaboration pattern involves the AI-Explorer generating initial hypothesis drafts, while the human-Evaluator performs the final scientific review. For instance, AI-Explorers generate candidate drug structures or material compositions, while human-Evaluators assess whether these designs are chemically feasible (Liu et al., 2024; Ni et al., 2024). Similarly, AI-Explorers scan vast chemical or protein spaces to find promising candidates. Human-Evaluators review the list to select the best options for testing (Kale et al., 2023; Swanson et al., 2025). In scientific idea generation, AI-Explorers combine concepts from the literature to propose new research directions or claims. Human-Evaluators validate these proposals against domain knowledge (Radensky et al., 2026).

AI-Explorer & Human-Controller. In interactive design tasks, the human-Controller guides the optimization process, while the AI-Explorer generates candidate solutions. In material and drug design, the human-Controller directs the design process by updating the requirements. The AI-Explorer then generates a new batch of molecules based on this guidance (Ansari et al., 2024; Ye et al., 2025). For drug discovery, the human-Controller guides the search toward a target protein and sets the property limits. The AI-Explorer generates and ranks candidate molecules to meet these goals (Kwon et al., 2024). Furthermore, in gene analysis, the human-Controller prioritizes which biological relationships are important for the search. The AI-Explorer can then use these relationship patterns to predict gene pairs that cause cell death (Jiang et al., 2025).

AI-Informer & Human-Explorer. The AI-Informer can gather evidence or predictions, while the human-Explorer analyzes them to propose new candidates. For example, in biomedical research, the AI-Informer integrates dispersed findings from the literature, helping the human-Explorer infer

potential hypotheses about relationships among biological factors (Jiang et al., 2026). Additionally, some AI-Informers can assist in retrieving relevant information from large volumes of data. For instance, this can enable human-Explorers to explore promising compound candidates (Shi et al., 2023b) or find similar image patches that support diagnostic theories (Corvò et al., 2021). In materials science, the AI-Informer forecasts physical properties such as conductivity, allowing the human-Explorer to combine these predictions to explore stable electrolytes for batteries (Pu et al., 2022).

Experiment Stage

The *experiment* stage involves designing and conducting tests to validate proposed hypotheses. In this phase, humans and AI collaborate to plan procedures and manage physical or computational processes to collect data.

AI-Controller & Human-Controller. During physical execution, the workflow functions as a shared control loop in which the AI and human manage different levels of the process. The AI-Controller handles immediate machine tasks, while the human-Controller directs the overall strategy. For instance, in robotic laboratories, the AI-Controller performs physical actions such as manipulating chemical samples. Human-Controller supervises the operation and updates targets based on real-time observations (Darvish et al., 2025). In materials laboratories, AI-Controller automates major steps from material preparation to characterization, while human-Controller provides oversight and adjusts actions based on the results (Ni et al., 2024). Similarly, in electron microscopy, the AI-Controller manages instrument settings to optimize data collection. Human-Controller monitors the live stream and directs the beam to explore relevant sample areas (Pratiush et al., 2024).

AI-Explorer & Human-Evaluator. When an experiment involves multiple steps, the AI-Explorer drafts a step-by-step plan. The human-Evaluator then reviews the final plan, corrects any errors, and determines whether it is usable. For example, in chemistry problem solving, the AI-Explorer uses specialized chemistry tools to generate candidate solutions and experimental procedures, while human-Evaluators assess the scientific correctness, completeness, and usability of the generated outputs (Bran et al., 2024). In gene editing, an AI-Explorer can generate an experimental plan, including suggested guide choices and key setup steps, with a human-Evaluator reviewing the final output before execution (Qu et al., 2025). For multi-step synthesis planning, the AI-Explorer proposes a full reaction route, and the human-Evaluator edits or replaces problematic steps in the route before laboratory work begins (Shi et al., 2023a). Furthermore, some AI-Explorers return a small set of candidate procedures along with a brief test plan, allowing human-Evaluators to edit the selected option and finalize what will be executed (Bazgir et al., 2025).

AI-Informer & Human-Explorer. Before conducting physical experiments, a common pattern is for the AI-Informer to provide fast predictions of possible outcomes, while the human-Explorer uses this information to search for promising experimental conditions and plans. For instance, in material design, the AI-Informer predicts how candidate 4D-printed mesh structures may deform, while the human-Explorer varies design parameters based on the AI prediction to identify configurations that achieve the desired shape changes (Yang et al., 2020). For biological simulations, the AI-Informer predicts yeast cell polarization under different parameter settings, allowing the human-Explorer to examine the predicted response landscape and select settings that match real-world observations (Hazarika et al., 2020). In chemical synthesis, the AI-Informer maps out available reaction pathways and risk estimates. Then the human-Explorer compares and adapts these options to develop a practical plan for laboratory execution (Wang et al., 2025a).

AI-Informer & Human-Controller. An emerging collaboration pattern pairs a human-Controller, who specifies and manipulates experimental conditions, with an AI-Informer, who participates in or simulates corresponding social interactions and makes the resulting behavioral dynamics observable. This pattern primarily appeared in recent social science research and supports the study of social interaction at multiple scales. In hybrid HAI team experiments, the human-Controller manipulates how team members communicate or coordinate, while the AI-Informer interacts with human participants and surfaces the resulting team behaviors (Yao et al., 2026). For interpersonal and small-group experiments, the human-Controller configures actor profiles, information access, and social goals. Guided by these specifications, the AI-Informer role-plays the designated actors and reveals how negotiation, information exchange, or group outcomes unfold (Sadiri Javadi et al., 2026). At the population level, the human-Controller changes the simulated user or introduces

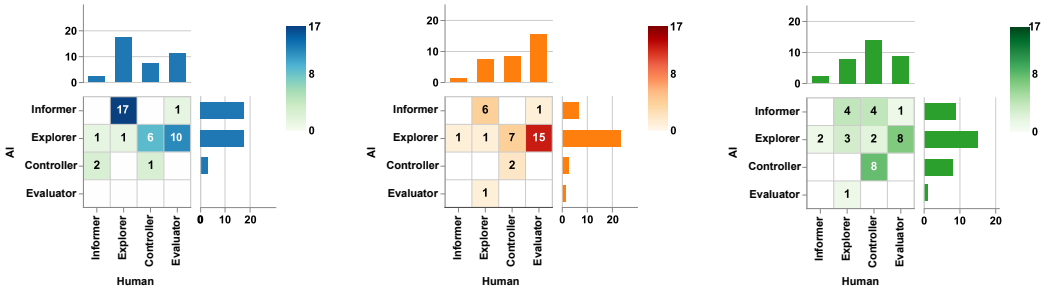


Figure 2: The HAI collaboration patterns across three scientific discovery stages.

external events. The AI-Informer then generates user behaviors and simulates how posts and social interactions evolve over time to support behavior analysis (Lin et al., 2025).

Role Differences Across Three Stages

Figure 2 illustrates a distinct imbalance in the distribution of human and AI roles across the three stages of scientific discovery. The *observation* stage frequently features the combination of AI-Informer and human-Explorer. In contrast, the *hypothesis* stage shows a significant shift toward the pairing of AI-Explorer and human-Evaluator. The *experiment* stage reveals a trend in which AI-Controller and human-Controller collaborate in executing protocols. This imbalance reflects the differing cognitive and operational demands across discovery stages. The *observation* stage is inherently open-ended, relying heavily on human insight to interpret phenomena and identify meaningful patterns, with AI primarily acting as an *Informer* that aggregates and summarizes data. As inquiry advances to the *hypothesis* stage, the problem space becomes more structured, allowing AI to systematically explore candidate hypotheses, while humans increasingly assume the *Evaluator* role to assess plausibility. In the *experiment* stage, the focus shifts to procedural execution, where requirements for correctness, safety, and reproducibility motivate a shared *Controller* role, with AI supporting automation and parameter control under human oversight.

These patterns indicate that HAI role allocation in scientific discovery dynamically reconfigures as epistemic uncertainty decreases and task structure increases. Humans dominate stages that demand sensemaking under ambiguity and normative judgment, whereas AI gains prominence as tasks become formalizable and computationally searchable. Notably, the *experiment* stage reflects a convergence rather than a transfer of control, highlighting the need to preserve meaningful human authority even in highly automated settings. This suggests that effective HAI collaboration should adapt role assignments across discovery stages, rather than imposing static responsibilities throughout the workflow.

4 Discussion

Built upon the key findings from our systematic survey, we discuss several challenges and potential future research directions for HAI collaboration in scientific discovery.

From Asymmetric Growth to Symbiotic Evolution. From Figure 3, we can observe that the evolution of roles reveals a complementary pattern: AI's development is specialized in computational tasks, while human involvement remains concentrated on high-judgment functions. Specifically, Figure 3a indicates that humans are least used as Informers but dominate as Evaluators and Controllers, with a strong presence as Explorers. This confirms that humans anchor the process in critical judgment, contextual reasoning, creative thinking, and ethical decision-making—areas requiring deep expertise and accountability. In contrast, Figure 3b shows that AI is predominantly deployed as an *Informer* and *Explorer*, with steady growth as a *Controller*. The *Evaluator* role remains minimal. This reflects a rational deployment of AI for its core strengths: processing data at scale, exploring solution spaces, and automating procedural workflows.

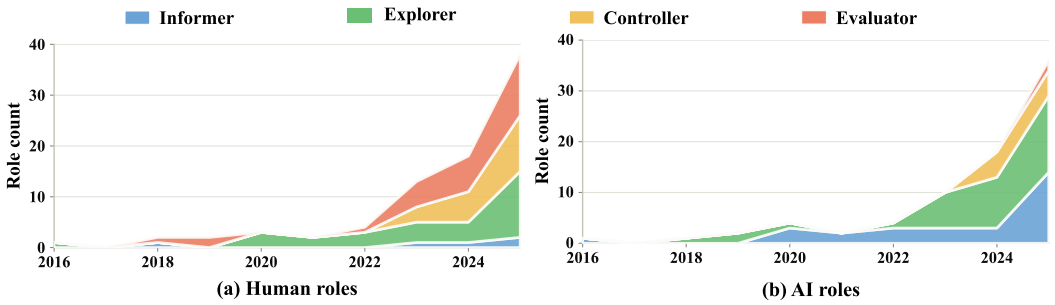


Figure 3: Evolution of human and AI roles. The year 2026 is excluded because we collected papers only up to June 2026.

The scarcity of the AI-Evaluator is the most pronounced example of this lag, stemming from fundamental technical gaps. Scientific evaluation requires: (1) Calibrated uncertainty quantification, whereas current models often provide overconfident point estimates, risking unreliable conclusions (Xie et al., 2025; Heo et al., 2025); (2) Causal and mechanistic reasoning, beyond surface pattern recognition (Shi et al., 2026); without understanding the underlying "why", evaluations may be misled by spurious correlations; (3) Contextual and normative judgment, aligning with tacit scientific standards–a challenge reflected in AI's difficulty with complex, value-laden trade-offs (Rezaei et al., 2025). These challenges in reliability, depth, and contextual alignment currently limit AI's role as a primary evaluator. Notably, the AI-Controller role also depends on robust reasoning and trustworthy autonomy. This shared need for reliability explains why its development, while progressing, remains more cautious than that of the well-established *Informer* and *Explorer* roles.

To evolve from the current functional division into a deeper, symbiotic partnership, future research must address these core limitations across roles. One potential direction is the development of AI-Controllers that go beyond basic workflow execution toward verifiable robustness. This necessitates benchmarks for failure recovery and methods for explainable workflow logic, ensuring systems can be audited and trusted in dynamic environments. As for AI-Evaluators, the focus should shift from automation to calibrated assistance. The immediate goal should be to develop tools that provide uncertainty-quantified assessments and evidence-attributed rationales, augmenting rather than replacing human judgment. In addition, effective collaborative design demands interfaces that formally position the human as a strategic supervisor. These systems should streamline the oversight of AI-generated options, making the human's role in guidance, interpretation, and final validation more efficient.

Generative Interfaces for Supporting Human Involvement. Most existing interfaces for HAI collaboration in scientific discovery are highly customized, often tailored to specific domains or discovery tasks. While such designs enable deep integration with domain workflows, they limit reusability and hinder generalization to other contexts. Recent advances in generative AI for user interface generation offer new opportunities to address this limitation (Chen et al., 2025). Rather than designing fixed, task-specific interfaces, future systems could dynamically generate interfaces that adapt to human responsibilities, such as exploration, evaluation, or control, throughout the scientific discovery process. Such context-aware interfaces have the potential to both enhance human oversight and maintain flexibility in exploration, paving the way for more effective HAI collaboration.

However, there are still several challenges. First, generative UI can implicitly constrain human exploration pathways. In scientific discovery, research directions emerge through iterative choices about variables, comparisons, and analytical operations rather than being predefined. When a UI is generated dynamically, these choices are partially delegated to the interface, which determines what controls, views, and exploration paths are available. Consequently, valid lines of inquiry may remain unexplored—not due to lack of scientific merit, but because they are interactionally unavailable. Therefore, it is critical to consider how to design generative interfaces that preserve open-ended exploration without inadvertently narrowing or biasing scientific discovery. Second, hallucinations in generative UIs pose heightened risks. Scientific data are often heterogeneous and multimodal, and even governed by domain-specific constraints, which can further exacerbate hallucinations in generative AI. Future work should therefore focus on developing domain-aware generative models,

benchmarks, and evaluation protocols that explicitly test whether generated interfaces respect data compatibility, experimental assumptions, and validity constraints.

Adaptive Role Assignments Between Human and AI. Existing HAI collaborative research typically predefines the roles of human and AI based on their abilities and limitations to solve scientific discovery tasks. However, scientific research is an inherently creative, exploratory, and non-linear process, in which research goals, hypotheses, and methods often evolve based on intermediate findings (Wang et al., 2023a). Such static designs fail to adapt to the uncertainty in the discovery process. Fixed role assignments may overlook scenarios where AI demonstrates unexpected proficiency in non-traditional tasks or where human intervention becomes necessary due to contextual judgments that exceed the predefined scope of automation. These limitations necessitate a paradigm shift toward dynamic and adaptive role assignment between humans and AI. Rather than framing HAI collaboration as a predefined division of labor, future research should conceptualize it as a problem of dynamic coordination. The roles of human and AI and their task allocation can be continuously negotiated based on contextual signals such as task difficulty, model confidence, experimental risk, and human cognitive load.

Empowering Embodied AI in Scientific Experiments. The papers in our corpus that address the *experiment* stage of scientific discovery primarily emphasize data analysis, while only a few investigate AI support for the practical execution of experimental processes. However, in domains like biology and chemistry, scientific discovery fundamentally relies on a large number of physical laboratory experiments rather than data analysis alone (Wang et al., 2023a). Embodied AI offers the potential to bridge this gap by converting AI models' planning capabilities into concrete experimental actions and operating directly at the laboratory bench (Pratiush et al., 2024). It tightens the coupling between scientific research decision-making and execution, allowing errors or uncertainties in AI-generated plans to directly affect physical experiments. This raises the stakes of human oversight and requires more continuous, real-time engagement across both conceptual and operational levels. These shifts highlight the need to carefully consider how humans can be more effectively integrated into the loop.

Long-term Implications for Leveraging AI in Scientific Discovery. As AI becomes increasingly powerful, the emergence of large language models has enabled systems to handle the entire research process, from hypothesis generation to experimental design and even paper writing (Schmidgall et al., 2025). In this context, it is increasingly important to consider the long-term implications of integrating AI into scientific workflows.

First, future work should establish standardized, auditable protocols for tracing AI involvement throughout the research process, rather than merely declaring AI usage. This can include the queries posed to AI systems, the full set of alternatives they generate, and the points at which human researchers intervene, modify, or reject AI suggestions. Such protocols enable process-level reproducibility, accountability, and responsible attribution of human and AI contributions, thereby supporting the long-term sustainability of the AI-augmented scientific ecosystem.

Second, the widespread use of AI in hypothesis exploration and decision-making raises important epistemic questions about how scientific reasoning may be reshaped over time. By mediating which hypotheses are generated, prioritized, or discarded, AI systems may systematically influence scientists' exploration strategies and cognitive trajectories. Future research should empirically identify the stages at which reliance on AI may gradually reshape key forms of human judgment (e.g., intuition, value-based reasoning, or cross-domain insight) and assess whether such shifts introduce systematic bias or ethical risk.

5 Conclusion

In this work, we presented a systematic review of HAI collaboration across the *observation*, *hypothesis*, and *experiment* stages of scientific discovery. By introducing a novel taxonomy of four roles of HAI collaboration (i.e., *Informer, Explorer, Evaluator*, and *Controller*), we provided a framework to better understand how AI and humans interact and contribute throughout the discovery process. Through our analysis, we identified key collaboration patterns and highlighted critical gaps related to role coordination, validation, and transparency. Finally, we outlined a research agenda for developing more adaptive, trustworthy, and efficient HAI systems.

AI USE STATEMENT

The authors used Kimi as a limited language-editing tool during the preparation of this manuscript. Its use was confined to refining the grammar, phrasing, and readability of text originally written by the authors. All suggested revisions were reviewed and, when necessary, modified by the authors, who assume full responsibility for the final manuscript.

REFERENCES

Hongran An and Zonglin Yang. MOOSE-copilot: A web-based interactive assistant for unified exploratory and fine-grained scientific hypothesis discovery. pp. 861–869, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-392-0. doi: 10.18653/v1/2026.acl-demo.85. URL https://aclanthology.org/2026.acl-demo.85/.

Mehrad Ansari, Jeffrey Watchorn, Carla E. Brown, and Joseph S Brown. dZiner: Rational Inverse Design of Materials with AI Agents. In *AI for Accelerated Materials Design - NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=4GU8XOvaAp.

Kavinda Athapaththu, Shiwei Chen, Yuan Fang, Sanchali Mitra, Yee Sin Ang, and Yong Wang. SemiConLens: Visual Analytics for 2D Semiconductor Discovery, 2026.

Adib Bazgir, Rama chandra Praneeth Madugula, and Yuwen Zhang. MatAgent: A human-in-the-loop multi-agent LLM framework for accelerating the material science discovery cycle. In *AI for Accelerated Materials Design - ICLR 2025*, 2025. URL https://openreview.net/forum?id=2Nm6Ef4tZD.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. doi: 10.1038/s42256-024-00832-8. URL https://www.nature.com/articles/s42256-024-00832-8.

Jiaqi Chen, Yanzhe Zhang, Yutong Zhang, Yijia Shao, and Diyi Yang. Generative interfaces for language models, 2025. URL https://arxiv.org/abs/2508.19227.

Zichen Cheng, Ziyue Lin, Yihang Yang, Zhongyu Wei, and Siming Chen. Interactive simulation and visual analysis of social media event dynamics with LLM-based multi-agent modeling. *Visual Informatics*, 9(3):100260, 2025. ISSN 2468-502X. doi: 10.1016/j.visinf.2025.100260. URL https://www.sciencedirect.com/science/article/pii/S2468502X25000439.

Isaac Cho, Wewnen Dou, Derek Xiaoyu Wang, Eric Sauda, and William Ribarsky. VAiRoma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):210–219, 2016. doi: 10.1109/TVCG.2015.2467971.

A. Corvò, H. S. Garcia Caballero, M. A. Westenberg, M. A. van Driel, and J. J. van Wijk. Visual Analytics for Hypothesis-Driven Exploration in Computational Pathology. *IEEE Transactions on Visualization and Computer Graphics*, 27(10):3851–3866, 2021. doi: 10.1109/TVCG.2020.2990336.

Kourosh Darvish, Marta Skreta, Yuchi Zhao, Naruki Yoshikawa, Sagnik Som, Miroslav Bogdanovic, Yang Cao, Han Hao, Haoping Xu, Alán Aspuru-Guzik, Animesh Garg, and Florian Shkurti. ORGANA: A robotic assistant for automated chemistry experimentation and characterization. *Matter*, 8(2):101897, 2025. ISSN 2590-2385. doi: 10.1016/j.matt.2024.10.015. URL https://www.sciencedirect.com/science/article/pii/S2590238524005423.

Mengjie Fan, Jinlu Yu, Daniel Weiskopf, Nan Cao, Huai-Yu Wang, and Liang Zhou. Visual Analysis of Multi-Outcome Causal Graphs. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):656–666, 2025. doi: 10.1109/TVCG.2024.3456346.

Carla Floricel, Nafiul Nipu, Mikayla Biggs, Andrew Wentzel, Guadalupe Canahuate, Lisanne Van Dijk, Abdallah Mohamed, C.David Fuller, and G.Elisabeta Marai. THALIS: Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):151–161, 2022. doi: 10.1109/TVCG.2021.3114810.

Carla Floricel, Andrew Wentzel, Abdallah Mohamed, C.David Fuller, Guadalupe Canahuate, and G.Elisabeta Marai. Roses Have Thorns: Understanding the Downside of Oncological Care Delivery Through Visual Analytics and Sequential Rule Mining. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1227–1237, 2024. doi: 10.1109/TVCG.2023.3326939.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an AI Co-Scientist, 2025. URL https://arxiv.org/abs/2502.18864.

Ricardo Luna Gutierrez, Sahand Ghorbanpour, Ejaz Rahman, Varchas Gopalaswamy, Riccardo Betti, Vineet Gundecha, Aarne Lees, and Soumyendu Sarkar. Human-in-the-Loop Meta Bayesian Optimization for Fusion Energy and Scientific Applications, 2026.

Subhashis Hazarika, Haoyu Li, Ko-Chih Wang, Han-Wei Shen, and Ching-Shan Chou. NNVA: Neural Network Assisted Visual Analysis of Yeast Cell Polarization Simulation. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):34–44, 2020. doi: 10.1109/TVCG.2019.2934591.

Juyeon Heo, Miao Xiong, Christina Heinze-Deml, and Jaya Narain. Do LLMs estimate uncertainty well in instruction-following? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=IHp3vOVQO2.

Jiayi Hong, Ross Maciejewski, Alain Trubuil, and Tobias Isenberg. Visualizing and Comparing Machine Learning Predictions to Improve Human-AI Teaming on the Example of Cell Lineage. *IEEE Transactions on Visualization and Computer Graphics*, 30(4):1956–1969, 2024. doi: 10.1109/TVCG.2023.3302308.

Chen Huang, Yang Deng, Wenqiang Lei, Jiancheng Lv, Tat-Seng Chua, and Jimmy Huang. How to Enable Effective Cooperation Between Humans and NLP Models: A Survey of Principles, Formalizations, and Beyond. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 466–488, 2025. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.22. URL https://aclanthology.org/2025.acl-long.22/.

Haejin Jeong, Hyoung-oh Jeong, Semin Lee, and Won-Ki Jeong. Optimal Dimensionality Selection Using Hull Heatmaps for Single-Cell Analysis. *Computer Graphics Forum*, 44(6):e70151, 2025. doi: 10.1111/cgf.70151. URL https://onlinelibrary.wiley.com/doi/10.1111/cgf.70151.

Haoran Jiang, Shaohan Shi, Shuhao Zhang, Jie Zheng, and Quan Li. SLInterpreter: An Exploratory and Iterative Human-AI Collaborative System for GNN-Based Synthetic Lethal Prediction. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):919–929, 2025. doi: 10.1109/TVCG.2024.3456325.

Haoran Jiang, Shaohan Shi, Yunjie Yao, Chang Jiang, and Quan Li. Hypochainer: A collaborative system combining llms and knowledge graphs for hypothesis-driven scientific discovery. *IEEE Transactions on Visualization and Computer Graphics*, 32(1):298–308, 2026. doi: 10.1109/TVCG.2025.3633887.

Zhihan Jiang, Handi Chen, Rui Zhou, Jing Deng, Xinchen Zhang, Running Zhao, Cong Xie, Yifang Wang, and Edith C.H. Ngai. HealthPrism: A Visual Analytics System for Exploring Children's Physical and Mental Health Profiles with Multimodal Data . *IEEE Transactions on Visualization & Computer Graphics*, 30(01):1205–1215, 2024. ISSN 1941-0506. doi: 10.1109/TVCG.2023.3326943. URL https://doi.ieeecomputersociety.org/10.1109/TVCG.2023.3326943.

Tabassum Kakar, Xiao Qin, Elke A. Rundensteiner, Lane Harrison, Sanjay K. Sahoo, and Suranjan De. DIVA: Exploration and Validation of Hypothesized Drug-Drug Interactions. *Computer Graphics Forum*, 38(3):95–106, 2019. doi: 10.1111/cgf.13674.

Bharat Kale, Austin Clyde, Maoyuan Sun, Arvind Ramanathan, Rick Stevens, and Michael E. Papka. ChemoGraph: Interactive Visual Exploration of the Chemical Space. *Computer Graphics Forum*, 42(3):13–24, 2023. doi: 10.1111/cgf.14807. URL https://www.osti.gov/biblio/1996240.

Bahavathy Kathirgamanathan, Tiago Janela, Elena Xerxa, Gennady Andrienko, Jürgen Bajorath, and Natalia Andrienko. From Prediction to Insight: Visual Analytics for Understanding Compound Potency Models. *IEEE Computer Graphics and Applications*, 46(3):133–140, 2026. doi: 10.1109/MCG.2026.3675766.

Yuya Kawakami, Daniel Cayan, Dongyu Liu, and Kwan-Liu Ma. ClimateSOM: a Visual Analysis Workflow for Climate Ensemble Datasets. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–11, 2025. doi: 10.1109/TVCG.2025.3634788.

Robert Krueger, Johanna Beyer, Won-Dong Jang, Nam Wook Kim, Artem Sokolov, Peter K. Sorger, and Hanspeter Pfister. Facetto: Combining Unsupervised and Supervised Learning for Hierarchical Phenotype Analysis in Multi-Channel Image Data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):227–237, 2020. doi: 10.1109/TVCG.2019.2934547.

Bum Chul Kwon, Simona Rabinovici-Cohen, Beldine Moturi, Ruth Mwaura, Kezia Wahome, Oliver Njeru, Miguel Shinyenyi, Catherine Wanjiru, Sekou Remy, William Ogallo, Itai Guez, Partha Suryanarayanan, Joseph Morrone, Shreyans Sethi, Seung-Gu Kang, Tien Huynh, Kenney Ng, Diwakar Mahajan, Hongyang Li, Matan Ninio, Shervin Ayati, Efrat Hexter, and Wendy Cornell. SPARK: harnessing human-centered workflows with biomedical foundation models for drug discovery. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/1015. URL https://doi.org/10.24963/ijcai.2024/1015.

Haotian Li, Yun Wang, and Huamin Qu. Where Are We So Far? Understanding Data Storytelling Tools from the Perspective of Human-AI Collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024. ISBN 9798400703300. doi: 10.1145/3613904.3642726. URL https://doi.org/10.1145/3613904.3642726.

Ziyue Lin, Yi Shan, Lin Gao, Xinghua Jia, and Siming Chen. SimSpark: Interactive Simulation of Social Media Behaviors. *Proc. ACM Hum.-Comput. Interact.*, 9(2), May 2025. doi: 10.1145/3711066. URL https://doi.org/10.1145/3711066.

Shengchao Liu, Jiongxiao Wang, Yijin Yang, Chengpeng Wang, Ling Liu, Hongyu Guo, and Chaowei Xiao. Conversational Drug Editing Using Retrieval and Domain Feedback. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=yRrPfKyJQ2.

Shuaihua Lu, Qionghua Zhou, Yixin Ouyang, Yilv Guo, Qiang Li, and Jinlan Wang. Accelerated Discovery of Stable Lead-Free Hybrid Organic-Inorganic Perovskites via Machine Learning. *Nature Communications*, 9(1):3405, 2018. doi: 10.1038/s41467-018-05761-w.

Deng Luo, Zainab Alsuwaykit, Dawar Khan, Ondřej Strnad, Tobias Isenberg, and Ivan Viola. DiffFit: Visually-Guided Differentiable Fitting of Molecule Structures to a Cryo-EM Map. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):558–568, 2025. doi: 10.1109/TVCG.2024.3456404.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, and Aixin Sun. SciAgent: Tool-augmented language models for scientific reasoning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15701–15736, November 2024. doi: 10.18653/v1/2024.emnlp-main.880. URL https://aclanthology.org/2024.emnlp-main.880/.

Vikram Mohanty, David Thames, Sneha Mehta, and Kurt Luther. Photo sleuth: combining human expertise and face recognition to identify historical portraits. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pp. 547–557, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362726. doi: 10.1145/3301275.3302301. URL https://doi.org/10.1145/3301275.3302301.

Eric Mörth, Kevin Sidak, Zoltan Maliga, Torsten Möller, Nils Gehlenborg, Peter Sorger, Hanspeter Pfister, Johanna Beyer, and Robert Krüger. Cell2Cell: Explorative Cell Interaction Analysis in Multi-Volumetric Tissue Data. *IEEE Transactions on Visualization and Computer Graphics*, 31(1): 569–579, 2025. doi: 10.1109/TVCG.2024.3456406.

Aman Nanda, Debjyoti Bhattacharya, Nayeon Kang, Shankar Kunwar, Shashank Priya, Jungho Ryu, Wesley Reinhart, Michael Lanagan, and Bed Poudel. Design and realization of high performance textured lead-free piezoelectric ceramics through human–AI collaboration. *Nature Communications*, 2026. doi: 10.1038/s41467-026-74568-x.

Ziqi Ni, Yahao Li, Kaijia Hu, Kunyuan Han, Ming Xu, Xingyu Chen, Fengqi Liu, Yicong Ye, and Shuxin Bai. MatPilot: An LLM-Enabled AI Materials Scientist Under the Framework of Human-Machine Collaboration, 2024. URL https://arxiv.org/abs/2411.08063.

Raúl Ortega and Jose Manuel Gomez-Perez. SciClaims: An end-to-end generative system for biomedical claim analysis. In Ivan Habernal, Peter Schulam, and Jörg Tiedemann (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 141–154, 2025. ISBN 979-8-89176-334-0. doi: 10.18653/v1/2025.emnlp-demos.11. URL https://aclanthology.org/2025.emnlp-demos.11/.

Reshika Palaniyappan Velumani, Meng Xia, Jun Han, Chaoli Wang, ALEXIS K LAU, and Huamin Qu. AQX: Explaining Air Quality Forecast for Verifying Domain Knowledge using Feature Importance Visualization. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pp. 720–733, 2022. ISBN 9781450391443. doi: 10.1145/3490099.3511150. URL https://doi.org/10.1145/3490099.3511150.

Ji Hwan Park, Vikash Prasad, Sydney Newsom, Fares Najar, and Rakhi Rajan. IdMotif: An Interactive Motif Identification in Protein Sequences. *IEEE Computer Graphics and Applications*, 44(3): 114–125, 2024. doi: 10.1109/MCG.2023.3345742.

Utkarsh Pratiush, Gerd Duscher, and Sergei Kalinin. Human-in-the-loop interface for Automated experiments in Electron Microscopy, Automated characterization. In *AI for Accelerated Materials Design - NeurIPS 2024*, 2024. URL https://openreview.net/forum?id=AB4EuiXkAr.

Jiansu Pu, Hui Shao, Boyang Gao, Zhengguo Zhu, Yanlin Zhu, Yunbo Rao, and Yong Xiang. matExplorer: Visual Exploration on Predicting Ionic Conductivity for Solid-state Electrolytes. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):65–75, 2022. doi: 10.1109/TVCG.2021.3114812.

Hangwei Qian, Yang He, Yaxin Shi, and Ivor Tsang. GenMatLab: a generative platform for inverse materials design. In *Proceedings of the Fortieth AAAI Conference on Artificial Intelligence and Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence and Sixteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'26/IAAI'26/EAAI'26. AAAI Press, 2026. ISBN 978-1-57735-906-7. doi: 10.1609/aaai.v40i48.42374. URL https://doi.org/10.1609/aaai.v40i48.42374.

Rui Qiu, Shijie Chen, Yu Su, Po-Yin Yen, and Han Wei Shen. Completing a systematic review in hours instead of months with interactive AI agents. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31559–31593, 2025. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1523. URL https://aclanthology.org/2025.acl-long.1523/.

Yuanhao Qu, Kaixuan Huang, Ming Yin, Kanghong Zhan, Dyllan Liu, Di Yin, Henry C. Cousins, William A. Johnson, Xiaotong Wang, Mihir Shah, et al. CRISPR-GPT for agentic automation of gene-editing experiments. *Nature Biomedical Engineering*, 2025. doi: 10.1038/s41551-025-01463-z. URL https://www.nature.com/articles/s41551-025-01463-z.

Marissa Radensky, Simra Shahid, Raymond Fok, Pao Siangliulue, Tom Hope, and Daniel S. Weld. Scideator: Human-llm compound system for scientific ideation through facet recombination and novelty evaluation. In *Proceedings of the ACM Conference on AI and Agentic Systems*,

CAIS '26, pp. 348–374, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400724152. doi: 10.1145/3786335.3813161. URL https://doi.org/10.1145/3786335.3813161.

Sebastian Ratzenböck, Verena Obermüller, Torsten Möller, João Alves, and Immanuel M. Bomze. Uncover: Toward Interpretable Models for Detecting New Star Cluster Members. *IEEE Transactions on Visualization and Computer Graphics*, 29(9):3855–3872, 2023. doi: 10.1109/TVCG.2022.3172560.

Chandan K. Reddy and Parshin Shojaee. Towards scientific discovery with generative AI: progress, opportunities, and challenges. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*. AAAI Press, 2025. ISBN 978-1-57735-897-8. doi: 10.1609/aaai.v39i27.35084. URL https://doi.org/10.1609/aaai.v39i27.35084.

MohammadHossein Rezaei, Yicheng Fu, Phil Cuvin, Caleb Ziems, Yanzhe Zhang, Hao Zhu, and Diyi Yang. EgoNormia: Benchmarking Physical-Social Norm Understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19256–19283, July 2025. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.985. URL https://aclanthology.org/2025.findings-acl.985/.

Yixiang Ruan, Chenyin Lu, Ning Xu, Yuchen He, Yixin Chen, Jian Zhang, Jun Xuan, Jianzhang Pan, Qun Fang, Hanyu Gao, et al. An automatic end-to-end chemical synthesis development platform powered by large language models. *Nature Communications*, 15(1):10160, 2024. doi: 10.1038/s41467-024-54457-x. URL https://doi.org/10.1038/s41467-024-54457-x.

María Virginia Sabando, Pavol Ulbrich, Matías Selzer, Jan Byška, Jan Mičan, Ignacio Ponzoni, Axel J. Soto, María Luján Ganuza, and Barbora Kozlíková. ChemVA: Interactive Visual Analysis of Chemical Compound Similarity in Virtual Screening. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):891–901, 2021. doi: 10.1109/TVCG.2020.3030438.

Vahid Sadiri Javadi, Fryderyk Róg, Aksa Aksa, Johanne Trippas, Svitlana Vakulenko, and Lucie Flek. CHARISMA: Character-Based Interaction Simulation with Multi-LLM Agents Toward Computational Social Psychology. In *Proceedings of the 2026 Conference on Human Information Interaction and Retrieval*, CHIIR '26, pp. 355–360, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400724145. doi: 10.1145/3786304.3787921. URL https://doi.org/10.1145/3786304.3787921.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using LLM agents as research assistants. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 5977–6043, November 2025. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.320. URL https://aclanthology.org/2025.findings-emnlp.320/.

Rui Sheng, Zelin Zang, Jiachen Wang, Yan Luo, Zixin Chen, Yan Zhou, Shaolun Ruan, and Huamin Qu. CellScout: Visual Analytics for Mining Biomarkers in Cell State Discovery. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–16, 2025. doi: 10.1109/TVCG.2025.3636102.

Chuhan Shi, Yicheng Hu, Shenan Wang, Shuai Ma, Chengbo Zheng, Xiaojuan Ma, and Qiong Luo. RetroLens: A Human-AI Collaborative System for Multi-step Retrosynthetic Route Planning. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023a. ISBN 9781450394215. doi: 10.1145/3544548.3581469. URL https://doi.org/10.1145/3544548.3581469.

Chuhan Shi, Fei Nie, Yicheng Hu, Yige Xu, Lei Chen, Xiaojuan Ma, and Qiong Luo. MedChemLens: An Interactive Visual Tool to Support Direction Selection in Interdisciplinary Experimental Research of Medicinal Chemistry. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):63–73, 2023b. doi: 10.1109/TVCG.2022.3209434.

Chuhan Shi, Xiaoquan Ren, Sicheng Song, Haobo Li, Rui Sheng, and Yushi Sun. Are LLMs ready for scientific discovery? a capability-oriented benchmark for AI scientists, 2026. URL https://arxiv.org/abs/2607.11079.

Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, 2025. doi: 10.1038/s41586-025-09442-9.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023a. doi: 10.1038/s41586-023-06221-2. URL https://www.nature.com/articles/s41586-023-06221-2.

Qianwen Wang, Kexin Huang, Payal Chandak, Marinka Zitnik, and Nils Gehlenborg. Extending the Nested Model for User-Centric XAI: A Design Study on GNN-based Drug Repurposing. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1266–1276, 2023b. doi: 10.1109/TVCG.2022.3209435.

Qipeng Wang, Rui Sheng, Shaolun Ruan, Xiaofu Jin, Chuhan Shi, and Min Zhu. SynthLens: Visual Analytics for Facilitating Multi-Step Synthetic Route Design. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):7647–7660, 2025a. doi: 10.1109/TVCG.2025.3552134.

Qipeng Wang, Shaolun Ruan, Rui Sheng, Yong Wang, Min Zhu, and Huamin Qu. Trajlens: Visual analysis for constructing cell developmental trajectories in cross-sample exploration. *IEEE Transactions on Visualization and Computer Graphics*, 32(1):1219–1229, 2026. doi: 10.1109/TVCG.2025.3634875.

Xingbo Wang, Samantha L. Huey, Rui Sheng, Saurabh Mehta, and Fei Wang. SciDaSynth: Interactive Structured Data Extraction From Scientific Literature With Large Language Model. *Campbell Systematic Reviews*, 21(1):e70073, 2025b. doi: 10.1002/cl2.70073. URL https://onlinelibrary.wiley.com/doi/10.1002/cl2.70073.

Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. GeneAgent: Self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22(8):1677–1685, 2025c. doi: 10.1038/s41592-025-02748-6.

Andrew Wentzel, Carla Floricel, Guadalupe Canahuate, Mohamed A. Naser, Abdallah S. Mohamed, Clifton D. Fuller, Lisanne van Dijk, and G. Elisabeta Marai. DASS Good: Explainable Data Mining of Spatial Cohort Data. *Computer Graphics Forum*, 42(3):283–295, 2023. doi: 10.1111/cgf.14830. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14830.

Austin P Wright, Peter Nemere, Adrian Galvin, Duen Horng Chau, and Scott Davidoff. Lessons from the development of an anomaly detection interface on the mars perseverance rover using the ishmap framework. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 91–105, 2023. ISBN 9798400701061. doi: 10.1145/3581641.3584036. URL https://doi.org/10.1145/3581641.3584036.

Qiujie Xie, Qingqiu Li, Zhuohao Yu, Yuejie Zhang, Yue Zhang, and Linyi Yang. An Empirical Analysis of Uncertainty in Large Language Model Evaluations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=J4xLuCt2kg.

Qi Xin, Quyu Kong, Hongyi Ji, Yue Shen, Yuqi Liu, Yan Sun, Zhilin Zhang, Zhaorong Li, Xunlong Xia, Bing Deng, and Yinqi Bai. BioInformatics Agent (BIA): Unleashing the Power of Large Language Models to Reshape Bioinformatics Workflow. bioRxiv preprint, 2024. URL https://doi.org/10.1101/2024.05.22.595240.

Chaoqing Xu, Tyson Neuroth, Takanori Fujiwara, Ronghua Liang, and Kwan-Liu Ma. A Predictive Visual Analytics System for Studying Neurodegenerative Disease Based on DTI Fiber Tracts. *IEEE Transactions on Visualization and Computer Graphics*, 29(4):2020–2035, 2023. doi: 10.1109/TVCG.2021.3137174.

Humphrey Yang, Kuanren Qian, Haolin Liu, Yuxuan Yu, Jianzhe Gu, Matthew McGehee, Yongjie Jessica Zhang, and Lining Yao. SimuLearn: Fast and Accurate Simulator to Support Morphing Materials Design and Workflows. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, pp. 71–84, 2020. ISBN 9781450375146. doi: 10.1145/3379337.3415867. URL https://doi.org/10.1145/3379337.3415867.

Bingsheng Yao, Jiaju Chen, Chaoran Chen, April Yi Wang, Toby Jia-Jun Li, and Dakuo Wang. Through the Lens of Human-Human Collaboration: A Configurable Research Platform for Exploring Human-Agent Collaboration. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3790879. URL https://doi.org/10.1145/3772318.3790879.

Jun-Hsiang Yao, Mingzheng Li, Jiayi Liu, Yuxiao Li, Jielin Feng, Jun Han, Qibao Zheng, Jianfeng Feng, and Siming Chen. DTBIA: An Immersive Visual Analytics System for Brain-Inspired Research. *IEEE Transactions on Visualization and Computer Graphics*, 31(6):3796–3808, 2025. doi: 10.1109/TVCG.2025.3567135.

Geyan Ye, Xibao Cai, Houtim Lai, Xing Wang, Junhong Huang, Longyue Wang, Wei Liu, and Xiangxiang Zeng. DrugAssist: a large language model for molecule optimization. *Briefings in Bioinformatics*, 26(1):bbae693, 2025. doi: 10.1093/bib/bbae693.

Wei Zhang, Jason K. Wong, Xumeng Wang, Youcheng Gong, Rongchen Zhu, Kai Liu, Zihan Yan, Siwei Tan, Huamin Qu, Siming Chen, and Wei Chen. CohortVA: A Visual Analytic System for Interactive Exploration of Cohorts based on Historical Data. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):756–766, 2023. doi: 10.1109/TVCG.2022.3209483.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From Automation to Autonomy: A Survey on Large Language Models in Scientific Discovery. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17744–17761, 2025. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.895. URL https://aclanthology.org/2025.emnlp-main.895/.

Xuhui Zhou, Zhe Su, Sophie Feng, Jiaxu Zhou, Jen-tse Huang, Hsien-Te Kao, Spencer Lynch, Svitlana Volkova, Tongshuang Wu, Anita Woolley, Hao Zhu, and Maarten Sap. SOTOPIA-s4: a user-friendly system for flexible, customizable, and large-scale social simulation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 350–360, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-191-9. doi: 10.18653/v1/2025.naacl-demo.30. URL https://aclanthology.org/2025.naacl-demo.30/.

APPENDIX

A SEARCH AND STUDY SELECTION PROTOCOL

A.1 SCOPE

The unit of inclusion is a paper describing a concrete interactive system or workflow for scientific discovery, published between January 2016 and June 2026. We adopt a broad view of scientific discovery that encompasses the collection and interpretation of empirical evidence, the formulation and evaluation of hypotheses, and the design or execution of experiments across diverse scientific domains. The scope therefore includes, but is not limited to, biomedical science, chemistry, materials science, climate and environmental science, physics, astronomy, social and behavioral science, and computational or empirical historical research, together with the interactive computational and visualization systems developed to support discovery in these domains.

A paper is eligible when its workflow contains at least one substantive human action and at least one substantive AI action that are directly coupled within the observation, hypothesis, or experiment stage of scientific discovery. An action is considered substantive when it materially shapes the scientific artifact under study, such as data, an observed pattern, a hypothesis, a candidate design, or an experimental procedure. Two actions are considered directly coupled when the output of one actor is taken up by the other within the same interaction episode. The mere co-existence of a human component and an AI component within a system, without a meaningful exchange between them, is not sufficient for inclusion.

This scope also excludes three common near-misses. First, we exclude generic writing, communication, or administrative assistance that does not directly contribute to an observation, hypothesis, or experiment. Second, we exclude fully automated pipelines in which the human merely initiates the workflow or inspects its final output, because such systems do not exhibit a substantive division of labor between humans and AI during scientific discovery. Third, we exclude data-labeling, annotation, and dataset-construction tools whose primary output is a corpus or dataset rather than a scientific observation, hypothesis, design, or experiment. Although these tools may support subsequent research, they operate before rather than within the scientific discovery process considered in this review.

A.2 Screening and Selection

Before full-text screening, we consolidated duplicate and superseded records. Different versions of the same work, including preprint, demonstration, conference, and journal versions, were treated as a single study. We first matched records using DOI information when available, then compared normalized titles, author lists, publication years, and system names. Ambiguous cases were resolved through manual inspection. When multiple versions described the same system or line of work, we retained the most complete eligible version, typically the journal article or the latest peer-reviewed publication. This conservative version-merging procedure prevented a single system from inflating the corpus through multiple publications.

After version consolidation, three co-authors independently read the full text of each remaining record and applied the three inclusion criteria. For every paper, the reviewers recorded an inclusion or exclusion decision together with the corresponding reason. Disagreements were resolved through discussion until consensus was reached. A paper entered the final corpus only when its full text provided evidence of at least one directly coupled human–AI collaboration episode. Specifically, an output produced by one actor had to be taken up by the other actor while working on the same scientific artifact, such as a data pattern, hypothesis, candidate design, or experimental procedure. The mere co-existence of human and AI components within the same system was not considered sufficient. Each qualifying collaboration episode subsequently served as the basis for identifying a human–AI role-pair instance during coding.

The complete selection flow was as follows. The keyword search across the venues listed in Table 1 returned 264 records. Iterative backward and forward snowballing over the retrieved papers and the two seed surveys added 32 records, resulting in 296 records in total. Version consolidation removed 87 duplicate or superseded records, leaving 209 distinct papers for full-text screening. The three co-authors independently reviewed all 209 papers. Applying the inclusion criteria excluded 144 papers and retained 65 papers in the final corpus. These 65 papers constitute the paper-level corpus reported in Appendix C; because a paper could contain multiple distinct collaboration episodes, it could subsequently yield more than one role-pair coding record.

A.3 Search Strategy

The primary method was a keyword search over the 2016–2026 period, spanning the main venues of the artificial intelligence, natural language processing, human–computer interaction, visualization, and science communities listed in Table 1. Each query paired a scientific discovery term for the relevant scientific domain with a human–AI collaboration term, following a keyword-based search strategy, with the three keyword groups listed in Table 2. We complemented the keyword search with iterative backward and forward snowballing over the references and citations of the retrieved papers and of two authoritative surveys on AI for science, continuing until no newly retrieved record satisfied the eligibility criteria.

Concretely, each query combined a scientific discovery or scientific domain term from Table 2 with a collaboration term. Representative composed queries include “protein” AND “human-in-the-loop”, “hypothesis generation” AND “LLM agent”, and “materials discovery” AND “human–AI collaboration”. At each venue we ran the product of the discovery terms and the collaboration terms and pooled the results before screening.

The two seed surveys, one on LLM autonomy in scientific discovery under a Tool/Analyst/Scientist taxonomy and the other on generative AI for scientific tasks, entered the process only as snowballing starting points rather than as sources of coded papers. We walked their reference lists and citation

Table 1: Venues searched, grouped by community.

Community	Venues
Artificial intelligence and machine learning	NeurIPS, ICML, ICLR, AAAI, IJCAI
Natural language processing	ACL, EMNLP, NAACL
Human–computer interaction	CHI, UIST, CSCW, CHIIR
Visualization	IEEE VIS, IEEE TVCG, IEEE Computer Graphics and Applications, EuroVis, Visual Informatics
Natural sciences	Nature and Nature Portfolio journals, Science

Table 2: Query construction. Each query paired a scientific-discovery term for the scientific domain with a collaboration term, following a keyword-based search strategy.

Keyword group	Representative terms
Scientific discovery	scientific discovery, observation, data exploration, pattern discovery, anomaly detection, hypothesis, hypothesis generation, hypothesis validation, experiment, experiment design, experimental execution, scientific reasoning
Scientific domain	protein structure and design, gene and genomics, molecule and drug discovery, materials science and design, chemistry, chemical synthesis and retrosynthesis, cell and structural biology, social interaction, behavioral science, computational social science, historical research, climate and environmental science, astronomy
Collaboration	human–AI collaboration, human-in-the-loop, interactive AI, LLM agent

graphs to reach the primary systems they cite and then applied the inclusion criteria below to each retrieved record. Backward snowballing followed the references of an included paper to earlier work, and forward snowballing followed its citations to later work. Because every newly included paper seeded a further pass, the procedure iterated to a fixed point, so that no eligible record reachable from an included paper was left unscreened.

B Codebook and Annotation Procedure

B.1 Role Codebook

The taxonomy assigns every coded action to one of four roles, and it defines each role by the function an actor performs on the shared scientific artifact rather than by whether the actor is a human or an AI. Every role therefore has both a human and an AI form, such as human-Explorer and AI-Explorer.

An *Informer* synthesizes, distills, retrieves, or articulates information, insights, or constraints from existing data, evidence, or intermediate analyses so that another actor can act. It reports on what already exists and neither commits to a new candidate nor rules on scientific merit. Typical Informer actions summarize a trajectory, visualize a feature-importance map, forecast a property, or annotate instrument state. The deciding test asks whether the output would still be useful as evidence even if no new candidate were retained. In THALIS, for instance, an AI-Informer summarizes longitudinal symptom trajectories for expert review.

An *Explorer* searches a space of data patterns, hypotheses, designs, routes, or experimental conditions and creates or selects a candidate that did not previously exist. Where the Informer only reports existing information, the Explorer commits to a new artifact. Typical Explorer actions generate a molecule, propose a hypothesis, or choose the next experimental condition. The deciding test asks whether the action created or selected an artifact that was not fixed beforehand. In ChemVA, for instance, a human-Explorer navigates a projected chemical space to identify molecular targets.

Table 3: Decision rules for frequently confused role boundaries.

Boundary	Decision rule	Split condition	Common error / example
Informer vs. Explorer	Information about an existing object is Informer. A newly selected or generated scientific candidate is Explorer. Classify by output artifact, not by model family.	Split only when retrieval/presentation and candidate generation are independently identifiable actions.	HypoChainer retrieval is Informer, and the generated hypothesis is Explorer. Do not call every prediction Informer or every output Explorer.
Explorer vs. Evaluator	Explorer proposes or searches candidates, and Evaluator rules on their scientific merit.	Split when one action generates candidates and another applies independent criteria to accept, reject, rank, or revise them.	RetroLens route generation is AI-Explorer, and chemist feasibility review is Human-Evaluator.
Evaluator vs. Controller	Evaluator judges an artifact, and Controller governs what the process does next or how it executes.	Split when an actor both validates content and separately changes objectives, sequence, runtime state, or instrument state.	Correcting a scientifically infeasible reaction step is evaluation. Authorizing and supervising a robot run is control.
Informer vs. Evaluator	Informer supplies evidence or explanation, and Evaluator issues a verdict or comparative merit judgment.	Split when evidence retrieval and explicit quality/veracity assessment are separately exposed.	An evidence panel is not automatically an AI-Evaluator.
Human-Informer vs. Human-Controller	Human-Informer contributes facts, annotations, targets, or constraints as information, while Human-Controller uses authority to steer progression or execution.	Split when annotations are supplied and the same person separately changes workflow state.	ISHMAP event annotations are information, and anomaly-search steering is control.
AI-Explorer vs. AI-Controller	Planning or proposing a protocol is Explorer. Controller requires runtime authority over tools, software workflow, or physical equipment.	Split when one component creates a procedure and another component runs it.	Experiment Designer is Explorer, and Hardware Executor is Controller. Internal tool use alone is not sufficient.

An *Evaluator* assesses the scientific merit of an artifact against evidence, criteria, or domain constraints and may accept, reject, rank, compare, or revise it. Where the Explorer proposes candidates, the Evaluator rules on them. Typical Evaluator actions verify feasibility, judge plausibility or novelty, and correct a scientifically flawed step. The deciding test asks whether the actor decides that an artifact is correct, credible, feasible, or worth keeping. In RetroLens, for instance, chemists act as human-Evaluators who assess AI-predicted synthetic routes for chemical feasibility.

A *Controller* governs how the workflow proceeds and how it executes at runtime by setting goals or constraints, sequencing steps, supervising tools or hardware, and recovering from failure. Where the Evaluator judges scientific merit, the Controller makes no such judgment and instead governs the process itself. Typical Controller actions steer a search, authorize a run, orchestrate a toolchain, and handle exceptions. The deciding test asks whether the actor decides how, when, or whether the workflow proceeds. In BIA, for instance, an AI-Controller orchestrates a bioinformatics toolchain and recovers from execution errors.

Two roles become hard to separate when one system both reports information and proposes candidates, or both judges an artifact and steers the workflow. Table 3 records, for each frequently confused pair, the decision rule, the condition under which one interaction is split into two coded patterns, and a worked example together with the mistake it prevents. A coder consults these rules whenever an action appears to fit two roles at once.

The four-role scheme is deliberately positioned between coarser and finer alternatives that we examined during coding. A coarser scheme would collapse the Informer and the Explorer into a single role by merging evidence organization with candidate generation. We rejected it because it obscures a basic difference in collaborative agency: an Informer makes existing information interpretable and actionable for another actor, whereas an Explorer actively expands the research space by proposing new patterns, hypotheses, or designs. A finer scheme would instead split the Controller into separate Planner and Executor roles. We rejected it because planning and execution were frequently carried out by the same actor, so the resulting categories depended heavily on system-specific implementation details and were too sparse to support comparison across papers. The four-role taxonomy therefore provides a stable and interpretable level of analysis for comparing human–AI collaboration across the corpus.

B.2 Stage Codebook

We place each coded interaction in one of the three stages of scientific discovery, following the decomposition adopted in the main paper. The scientific function of the shared artifact fixes the stage, not the technique that produces it, so the same method can serve different stages in different systems. Table 4 states the operational definition of each stage, its typical artifacts, and the boundary test that separates it from the adjacent stages.

B.3 Annotation Procedure

With the codebook fixed, the coders applied it to the corpus one paper at a time, following the six steps below. Each record they produced is a role-pair instance: one AI action and one human action that work on the same artifact within a single discovery stage, with one role assigned to each.

1. **Find a collaboration episode.** The coder read the paper and located a concrete point at which a human and an AI act on the same object, such as the AI proposing a synthetic route and a chemist then judging its feasibility. This coupled episode, not the system as a whole, was the unit of coding, and a paper that described several distinct episodes yielded one record for each.
2. **Decide the stage.** The coder placed the episode in the observation, hypothesis, or experiment stage according to what the shared object is. A pattern or view drawn from data is observation, a proposed explanation or design is hypothesis, and a plan or execution of a test is experiment. Table 4 gives the test the coder used when an object could belong to two stages.
3. **Label the two roles.** The coder named the object the two actors exchanged and then gave the AI action and the human action one role each from the role codebook, judging each

Table 4: Operational stage definitions.

Stage	Operational definition	Typical artifact	Boundary test
Observation	Collecting, structuring, or examining data or phenomena to identify patterns, anomalies, or questions requiring explanation.	Data view, cluster, anomaly, trajectory, extracted evidence, intermediate result.	If the central artifact is a proposed explanation or design, use Hypothesis. If it is an executable test or runtime action, use Experiment.
Hypothesis	Formulating, searching, comparing, or evaluating tentative explanations, claims, molecular/material designs, or other testable candidates.	Hypothesis, scientific claim, candidate molecule/material, causal relation, conceptual design.	A predicted class label alone is not necessarily a hypothesis. The artifact must function as a candidate explanation or design.
Experiment	Designing, planning, simulating, executing, or supervising a test intended to validate a hypothesis or collect decisive evidence.	Protocol, reaction route, guide, parameter configuration, test plan, robot/instrument state.	Conceptual candidate generation with no testing procedure remains Hypothesis. Evidence display with no test remains Observation.

action by what it does to that object rather than by the name of the model or the system. The two labels were assigned separately, so that reading the AI as an Explorer did not by itself make the human an Evaluator.

4. **Check the coupling.** The coder kept the pair only when one actor's output was actually taken up by the other in the same episode, and dropped cases where a human and an AI shared a system without exchanging anything. A paper with two genuinely different episodes produced two records, but a system with several roles was never expanded into every possible human and AI combination.
5. **Record the evidence.** For each labeled action the coder wrote a one-line reason and the exact location in the paper, such as a section or figure, and marked the record as confident or borderline so that unclear cases could be revisited.
6. **Reconcile the records.** Because three of the six co-authors coded each paper independently, they then compared their records, resolved every disagreement through discussion to consensus, and deferred to a senior author when consensus could not be reached.

Two kinds of activity were never counted. Internal AI behavior, such as agents debating each other or a model checking its own output, counted only when it produced a visible action that a human then engaged with. Contributions made only while building the system, such as interviews or design feedback, were treated as development rather than as collaboration during discovery.

C Coding Results

Table 5 lists every included paper with its discovery stage(s) and its human and AI roles.

Table 5: Complete 65-paper corpus with discovery stages and roles.

Paper	Stage	Human role	AI role
A Predictive Visual Analytics System for Studying Neurodegenerative Disease Based on DTI Fiber Tracts (Xu et al., 2023)	Observation	Explorer	Informer

Paper	Stage	Human role	AI role
BioInformatics Agent (BIA): Unleashing the Power of Large Language Models to Reshape Bioinformatics Workflow (Xin et al., 2024)	Observation	Controller	Explorer, Controller
CellScout: Visual Analytics for Mining Biomarkers in Cell State Discovery (Sheng et al., 2025)	Observation	Explorer, Evaluator	Informer, Explorer
ChemVA: Interactive Visual Analysis of Chemical Compound Similarity in Virtual Screening (Sabando et al., 2021)	Observation	Explorer	Informer
ClimateSOM: a Visual Analysis Workflow for Climate Ensemble Datasets (Kawakami et al., 2025)	Observation	Explorer	Informer
CohortVA: A Visual Analytic System for Interactive Exploration of Cohorts based on Historical Data (Zhang et al., 2023)	Observation	Evaluator	Explorer
Completing A Systematic Review in Hours instead of Months with Interactive AI Agents (Qiu et al., 2025)	Observation	Controller	Explorer
DASS Good: Explainable Data Mining of Spatial Cohort Data (Wentzel et al., 2023)	Observation	Controller, Evaluator	Informer, Explorer
DiffFit: Visually-Guided Differentiable Fitting of Molecule Structures to a Cryo-EM Map (Luo et al., 2025)	Observation	Evaluator	Explorer
DTBIA: An Immersive Visual Analytics System for Brain-Inspired Research (Yao et al., 2025)	Observation	Explorer	Informer
AQX: Explaining Air Quality Forecast for Verifying Domain Knowledge using Feature Importance Visualization (Palaniyappan Velumani et al., 2022)	Observation	Explorer	Informer
Extending the Nested Model for User-Centric XAI: A Design Study on GNN-based Drug Repurposing (Wang et al., 2023b)	Observation	Evaluator	Explorer
Facetto: Combining Unsupervised and Supervised Learning for Hierarchical Phenotype Analysis in Multi-Channel Image Data (Krueger et al., 2020)	Observation	Explorer	Informer
GeneAgent: Self-Verification Language Agent for Gene-Set Analysis Using Domain Databases (Wang et al., 2025c)	Observation	Evaluator	Explorer
HealthPrism: A Visual Analytics System for Exploring Children's Physical and Mental Health Profiles with Multimodal Data (Jiang et al., 2024)	Observation	Explorer	Informer
IdMotif: An Interactive Motif Identification in Protein Sequences (Park et al., 2024)	Observation	Explorer	Informer
Lessons from the Development of an Anomaly Detection Interface on the Mars Perseverance Rover using the ISHMAP Framework (Wright et al., 2023)	Observation	Informer, Controller	Explorer
Optimal Dimensionality Selection Using Hull Heatmaps for Single-Cell Analysis (Jeong et al., 2025)	Observation	Explorer	Informer
Roses Have Thorns: Understanding the Downside of Oncological Care Delivery Through Visual Analytics and Sequential Rule Mining (Floricel et al., 2024)	Observation	Evaluator	Explorer
THALIS: Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy (Floricel et al., 2022)	Observation	Explorer, Evaluator	Informer, Explorer

Paper	Stage	Human role	AI role
Visualizing and Comparing Machine Learning Predictions to Improve Human-AI Teaming on the Example of Cell Lineage (Hong et al., 2024)	Observation	Evaluator	Explorer
Uncover: Toward Interpretable Models for Detecting New Star Cluster Members (Ratzenböck et al., 2023)	Observation	Controller	Explorer
Cell2Cell: Explorative Cell Interaction Analysis in Multi-Volumetric Tissue Data (Mörth et al., 2025)	Observation	Explorer	Informer
TrajLens: Visual Analysis for Constructing Cell Developmental Trajectories in Cross-Sample Exploration (Wang et al., 2026)	Observation	Explorer	Informer
Visual Analysis of Multi-Outcome Causal Graphs (Fan et al., 2025)	Observation	Controller	Explorer
VAiRoma: A Visual Analytics System for Making Sense of Places, Times, and Events in Roman History (Cho et al., 2016)	Observation	Explorer	Informer
Interactive simulation and visual analysis of social media event dynamics with LLM-based multi-agent modeling (Cheng et al., 2025)	Observation	Explorer	Informer
HypoChainer: a Collaborative System Combining LLMs and Knowledge Graphs for Hypothesis-Driven Scientific Discovery (Jiang et al., 2026)	Hypothesis	Explorer, Evaluator	Informer, Explorer
Accelerated Discovery of Stable Lead-Free Hybrid Organic-Inorganic Perovskites via Machine Learning (Lu et al., 2018)	Hypothesis	Informer, Evaluator	Explorer
Conversational Drug Editing Using Retrieval and Domain Feedback (Liu et al., 2024)	Hypothesis	Evaluator	Explorer
ChemoGraph: Interactive Visual Exploration of the Chemical Space (Kale et al., 2023)	Hypothesis	Explorer, Evaluator	Explorer
DIVA: Exploration and Validation of Hypothesized Drug-Drug Interactions (Kakar et al., 2019)	Hypothesis	Evaluator	Explorer
DrugAssist: a large language model for molecule optimization (Ye et al., 2025)	Hypothesis	Controller, Evaluator	Explorer
dZiner: Rational Inverse Design of Materials with AI Agents (Ansari et al., 2024)	Hypothesis	Controller	Explorer, Controller
matExplorer: Visual Exploration on Predicting Ionic Conductivity for Solid-state Electrolytes (Pu et al., 2022)	Hypothesis	Explorer	Informer
MedChemLens: An Interactive Visual Tool to Support Direction Selection in Interdisciplinary Experimental Research of Medicinal Chemistry (Shi et al., 2023b)	Hypothesis	Explorer	Informer
MOOSE-Copilot: A Web-Based Interactive Assistant for Unified Exploratory and Fine-Grained Scientific Hypothesis Discovery (An & Yang, 2026)	Hypothesis	Controller	Explorer
SemiConLens: Visual Analytics for 2D Semiconductor Discovery (Athapaththu et al., 2026)	Hypothesis	Explorer	Informer
GenMatLab: A Generative Platform for Inverse Materials Design (Qian et al., 2026)	Hypothesis	Controller, Evaluator	Explorer
From Prediction to Insight: Visual Analytics for Understanding Compound Potency Models (Kathirgamanathan et al., 2026)	Hypothesis	Explorer	Informer
Design and Realization of High Performance Textured Lead-Free Piezoelectric Ceramics through Human-AI Collaboration (Nanda et al., 2026)	Hypothesis	Controller, Evaluator	Explorer

Paper	Stage	Human role	AI role
SciClaims: An End-to-End Generative System for Biomedical Claim Analysis (Ortega & Gomez-Perez, 2025)	Observation, Hypothesis, Experiment	Evaluator	Informer, Explorer
Scideator: Human-LLM Scientific Idea Generation Grounded in Research-Paper Facet Recombination (Radensky et al., 2026)	Hypothesis	Explorer, Evaluator	Explorer, Evaluator
SLInterpreter: An Exploratory and Iterative Human-AI Collaborative System for GNN-Based Synthetic Lethal Prediction (Jiang et al., 2025)	Hypothesis	Controller	Explorer
SPARK: harnessing human-centered workflows with biomedical foundation models for drug discovery (Kwon et al., 2024)	Hypothesis	Controller, Evaluator	Explorer
Visual Analytics for Hypothesis-Driven Exploration in Computational Pathology (Corvò et al., 2021)	Hypothesis	Explorer	Informer
Photo Sleuth: Combining Human Expertise and Face Recognition to Identify Historical Portraits (Mohanty et al., 2019)	Hypothesis	Evaluator	Explorer
Human-in-the-loop interface for Automated experiments in Electron Microscopy, Automated characterization (Pratiush et al., 2024)	Experiment	Explorer, Controller	Explorer, Controller
SimuLearn: Fast and Accurate Simulator to Support Morphing Materials Design and Workflows (Yang et al., 2020)	Experiment	Explorer	Informer
NNVA: Neural Network Assisted Visual Analysis of Yeast Cell Polarization Simulation (Hazarika et al., 2020)	Experiment	Explorer	Informer, Explorer
SynthLens: Visual Analytics for Facilitating Multi-Step Synthetic Route Design (Wang et al., 2025a)	Experiment	Explorer, Evaluator	Informer
An automatic end-to-end chemical synthesis development platform powered by large language models (Ruan et al., 2024)	Observation, Hypothesis, Experiment	Informer, Controller, Evaluator	Explorer, Controller
RetroLens: A Human-AI Collaborative System for Multi-step Retrosynthetic Route Planning (Shi et al., 2023a)	Experiment	Explorer, Evaluator	Explorer
CRISPR-GPT for agentic automation of gene-editing experiments (Qu et al., 2025)	Experiment	Informer, Controller, Evaluator	Explorer, Controller
Augmenting large language models with chemistry tools (Bran et al., 2024)	Experiment	Evaluator	Explorer
ORGANA: A robotic assistant for automated chemistry experimentation and characterization (Darvish et al., 2025)	Experiment	Controller	Explorer, Controller
Human-in-the-Loop Meta Bayesian Optimization for Fusion Energy and Scientific Applications (Gutierrez et al., 2026)	Experiment	Informer, Evaluator	Explorer
Through the Lens of Human-Human Collaboration: A Configurable Research Platform for Exploring Human-Agent Collaboration (Yao et al., 2026)	Experiment	Controller	Informer
SOTOPIA-S4: a user-friendly system for flexible, customizable, and large-scale social simulation (Zhou et al., 2025)	Experiment	Controller	Informer
SimSpark: Interactive Simulation of Social Media Behaviors (Lin et al., 2025)	Experiment	Controller	Informer
CHARISMA: Character-Based Interaction Simulation with Multi-LLM Agents Toward Computational Social Psychology (Sadiri Javadi et al., 2026)	Experiment	Controller	Informer
MatAgent: A human-in-the-loop multi-agent LLM framework for accelerating the material science discovery cycle (Bazgir et al., 2025)	Observation, Hypothesis, Experiment	Informer, Explorer, Controller, Evaluator	Informer, Explorer, Controller

Paper	Stage	Human role	AI role
The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies (Swanson et al., 2025)	Observation, Hypothesis, Experiment	Explorer, Controller, Evaluator	Explorer, Controller, Evaluator
Towards an AI Co-Scientist (Gottweis et al., 2025)	Observation, Hypothesis, Experiment	Explorer, Controller, Evaluator	Informer, Explorer, Controller
MatPilot: An LLM-Enabled AI Materials Scientist Under the Framework of Human-Machine Collaboration (Ni et al., 2024)	Observation, Hypothesis, Experiment	Explorer, Controller, Evaluator	Informer, Explorer, Controller