Review



# A Survey of Human-AI Collaboration for Scientific Discovery

Chuhan Shi , Xiaoquan Ren , Yifang Wang , Junze Li , Yushi Sun , Yawen Luo , Rui Sheng *

Posted Date: 5 February 2026

doi: 10.20944/preprints202601.0405.v2

Keywords: human-AI collaboration

 doi:10.20944/preprints202601.0405.v2



*Review*

# A Survey of Human-AI Collaboration for Scientific Discovery

**Chuhan Shi [1], Xiaoquan Ren [1], Yifang Wang [2], Junze Li [3], Yushi Sun [3], Yawen Luo [3] and Rui Sheng [3,*]**

[1] Southeast University
[2] Florida State University
[3] Hong Kong University of Science and Technology
* Correspondence: rshengac@connect.ust.hk

**Abstract**

Artificial intelligence (AI) is increasingly integrated into scientific discovery processes, such as protein design, gene analysis, and materials research, significantly enhancing the efficiency of discoveries in these fields. While much 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 rapid growth in human-in-the-loop and collaborative systems, the field lacks a unifying survey that explains how humans and AI actually collaborate across the scientific discovery life-cycle. In this paper, we present a systematic review of human–AI (HAI) collaboration for scientific discovery. Specifically, we have identified four representative roles of humans and AI. 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 human–AI systems in scientific discovery.

**Keywords:** human-AI collaboration

## 1. Introduction

The integration of artificial intelligence (AI) into scientific research has progressed from a methodological enhancement to a fundamental paradigm shift [1,2]. Early AI models primarily functioned as computational tools, facilitating low-level analytical tasks, such as pattern extraction and representation learning [3,4]. More recently, advances in AI, particularly large language models (LLMs), have introduced stronger capabilities for reasoning [5,6], exerting a transformative effect on the scientific discovery process. For example, some AI systems even demonstrate the ability to plan and execute experiments autonomously [7–9].

 doi:10.20944/preprints202601.0405.v2



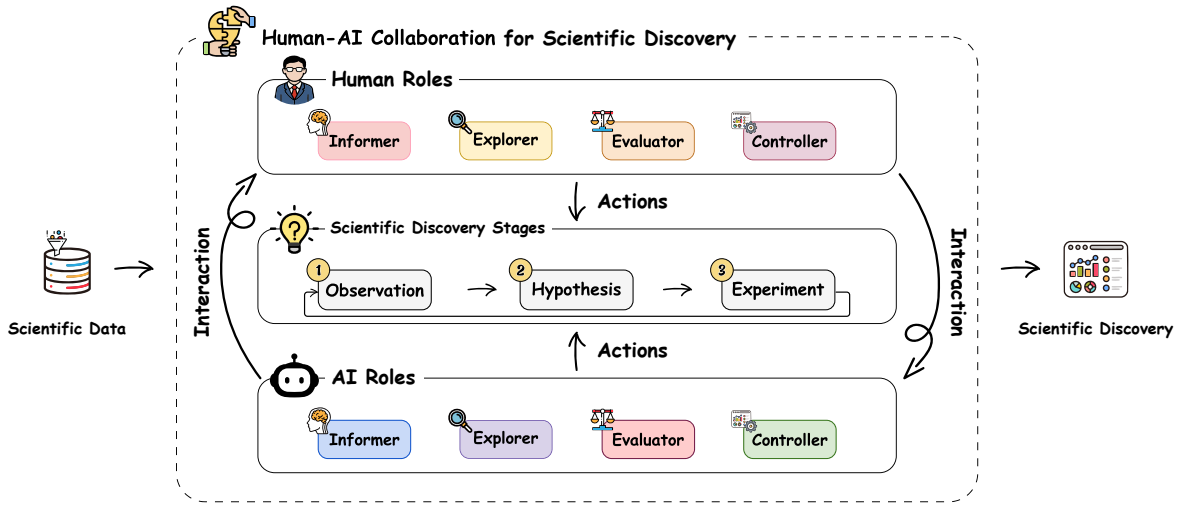


**Figure 1.** Our taxonomy characterizes research on human-AI collaborative scientific discovery from four roles of human and AI across the three stages of scientific discovery.

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. [9] reviews LLM-based systems for scientific discovery and proposes a three-level autonomy taxonomy (i.e., Tool, Analyst, Scientist). Reddy and Shojaee [10] provides a survey of generative AI for scientific tasks and summarizes challenges in building AI systems for scientific discovery. Consequently, our theoretical understanding of how humans and AI can effectively collaborate together throughout the scientific research process remains limited. Although surveys on human-NLP cooperation [11] and general human-AI interaction discuss relevant interaction principles [12–14], they often overlook the specific context of scientific discovery. Therefore, there is a lack of a structured framework for understanding the human-AI partnership in scientific discovery.

To bridge this gap, we presented a systematic taxonomy for human-AI collaboration in scientific discovery. First, we identified four roles of human and AI based on a systematic review of 51 papers and anchored our analysis in the established three stages of scientific discovery [15] (i.e., observation, hypothesis, and experiment). This established a unified framework that organized collaborative dynamics into consistent and comparable units. 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 differences across the three stages. We conclude by outlining five open challenges and future directions for establishing efficient and trustworthy human–AI partnerships. The main contributions of this paper are summarized as follows:

- We presented a comprehensive survey that systematically reviews human-AI collaboration specifically in scientific discovery.
- We introduced a novel taxonomy that defines the 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 human-AI partnerships in the scientific discovery process.

## 2. Methodology

In this section, we outline the methodology for collecting a corpus of 51 papers on human-AI collaboration for scientific discovery and the coding process used to identify the roles of human and AI.

 doi:10.20944/preprints202601.0405.v2



### 2.1. Paper Collection

To assemble a high-quality corpus focused on human-AI collaboration for scientific discovery, we implemented a systematic selection process encompassing research published between 2015 and 2025. We began by identifying seed papers from authoritative surveys on AI for science [9,10]. Using these papers as a baseline, we then conducted an iterative snowballing procedure, examining references and citations to identify relevant work until no further relevant studies emerged.

To ensure relevance, we applied several screening criteria to select papers for inclusion in our corpus. We began by reviewing the abstracts, and if necessary, examined other sections. Each paper had to present an interactive system or workflow explicitly designed to facilitate scientific discovery. As a result, we excluded papers whose contribution was the development of a fully automated algorithm. Additionally, we excluded studies focused solely on data labeling tasks (e.g., SciDaSynth [16]), even if the paper suggested that the dataset could contribute to future scientific discovery, as these works were considered too preliminary. To maintain high quality, we included only published papers or preprints with more than 100 citations. The final corpus consisted of 51 papers from a variety of reputable venues, such as Nature, ACL, EMNLP, CHI, and TVCG.

### 2.2. Paper Coding

Initially, six co-authors independently coded a subset of the corpus to derive the roles of humans and AI. Through weekly discussions, they resolved conflicts and unified the coding results, ultimately identifying four distinct roles. To better analyze their functions in scientific discovery, we followed a commonly used three-stage decomposition of the scientific process (i.e., observation, hypothesis, experiment) [15], coding each paper according to which stage it belonged to. 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 hypotheses—tentative, testable explanations or predictions that guide inquiry. The experiment stage then designs and conducts controlled studies or analyses to test these hypotheses, using the results to validate, refine, or reject them, often leading to new observations and continuing the discovery cycle.

## 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 human–AI 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).

◇ **Informer**. The Informer synthesizes, distills, or articulates key information, insights, or constraints from raw data or intermediate analyses to guide the actions of other roles. For instance, the AI-Informer in THALIS extracts temporal patterns from longitudinal symptom records in cancer therapy [17]. This provides a summarized trajectory view for experts to analyze patient responses to treatment. Similarly, in ISHMAP for Mars rover operations [18], 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. Compared with Informer, whose output provides low-level data insights, the Explorer directly generates candidates tailored to the specific needs of each stage in the scientific discovery process. For instance, in the hAE interface [7], the

 doi:10.20944/preprints202601.0405.v2



AI-Explorer searches the parameter space to select the next experimental conditions for electron microscopy. ChemVA [19] enables the human-Explorer to interactively navigate a projected chemical space to identify molecular targets.

◇ **Evaluator**. Once artifacts are proposed, their scientific merit must be rigorously evaluated and even refined. The Evaluator assesses observed patterns, hypotheses, or experimental designs based on predefined criteria, evidence, and domain constraints, revising them as necessary to meet quality standards. For instance, in RetroLens [20], chemists serve as the human-Evaluator. Specifically, they can assess AI-predicted synthetic routes for chemical feasibility or refine the synthetic steps by themselves.

◇ **Controller**. The Controller oversees the scientific discovery workflow to ensure correct and constraint-compliant execution, intervening when necessary to adjust procedures and handle runtime exceptions. This role is central to BIA [21], where the AI-Controller orchestrates the execution of complex bioinformatics toolchains, dynamically handling errors and modifying the workflow logic to ensure successful completion.

### *3.2. Common Collaboration Patterns Within Each Stage of Scientific Discovery*

In this section, we introduce common human–AI 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 human-AI collaboration and derive collaboration patterns from them to ensure that our analysis reflects meaningful human–AI collaboration rather than mere role co-existence

#### 3.2.1. 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 human-AI collaboration pattern in the observation stage involves an AI-Informer transforming raw data into structured representations (e.g., embeddings or feature importance maps), supporting a human-Explorer to efficiently identify clusters, trends, and outliers in scientific data. For example, some AI-Informers map high-dimensional data into a two-dimensional view, allowing human-Explorers to observe how patterns change across different conditions [22,23]. In biological domains, AI-Informers visualize tissue interactions [24], pediatric health profiles [25], neural connections [26], and cell trajectories [27]. Human-Explorers follow these visual pathways to uncover disease trends or developmental stages. Other tools organize items by similarity. The AI-Informers can also cluster compounds or highlight sequence motifs, allowing human-Explorers to search for promising chemicals [19], biomarkers [28], or protein functions [29]. Furthermore, the AI-Informers can rank the influence of features on predictions. The Human-Explorers investigate these cues to understand air pollution drivers [30], explore raw fiber tracts [31], and analyze phenotype images [32].

*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 highlights connections among different drugs for repurposing, enabling the human Evaluator to assess and verify the underlying biological rationale [33]. In structural biology, the AI-Explorer can generate 3D atomic models to match density maps, allowing the human-Evaluator to examine the structural configuration [34]. 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 [17] or hospital protocols [35]. 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

 doi:10.20944/preprints202601.0405.v2



consensus [36]. Additionally, the AI-Explorer can retrieve and explore the functional roles of gene groups, while the human Evaluator reviews this information to validate their biological relevance [37].

*AI-Explorer* & *Human-Controller*. With autonomous tools, observation follows an iterative search loop. The human-Controller directs the search process, while the AI-Explorer scans the data. For literature discovery, the human-Controllers guide the search direction, while the AI-Explorers navigate databases to summarize relevant papers [21,38]. To identify star clusters or spatial groups, the human-Controller adjust the detection parameters. The AI-Explorer scans astronomical surveys to locate new star clusters [39] or analyzes spatial population data to find notable spatial groups [40]. In causal analysis, the human-Controller refines the search constraints, enabling the AI-Explorer to explore cause-and-effect relationships [41]. Additionally, for Mars rover operations, the human-Controller sets anomaly criteria to enable the AI-Explorer to detect signal anomalies [18].

#### 3.2.2. 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 the initial hypothesis draft, 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 [42–44]. 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 [45,46]. 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 [47–49].

*AI-Explorer* & *Human-Controller*. In interactive design tasks, the human-Controller guides the optimization process, while the AI-Explorer generates candidate solutions. For example, 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 [50,51]. 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 [52]. 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 [53].

*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 [54]. 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 [55] or find similar image patches that support diagnostic theories [56]. 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 [57].

#### 3.2.3. 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

 doi:10.20944/preprints202601.0405.v2



chemical samples. Human-Controller supervises the operation and updates targets based on real-time observations [58]. 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 [44]. 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 [7].

*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, some AI-Explorers draft chemistry workflows by calling external chemistry tools during the planning phase, then return a complete procedure for human-Evaluators to review [59,60]. 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 [61]. 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 [20]. 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 [46,62,63].

*AI-Informer* & *Human-Explorer*. Before conducting physical experiments, the AI-Informer quickly forecasts potential results. The human-Explorer navigates these predictions to find the optimal experimental conditions. For instance, in material design, the AI-Informer predicts how structures deform, while the human-Explorer searches the design space to drive configurations that achieve the desired shape changes [64]. For biological simulations, the AI-Informer predicts yeast cell polarization, allowing the human-Explorer to navigate the parameter space and explore settings that match real-world observations [65]. In chemical synthesis, the AI-Informer maps out alternative reaction pathways and provides risk estimates. Then the human-Explorer explores these pathways to develop a practical plan for laboratory execution [66].

### 3.2.4. 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.

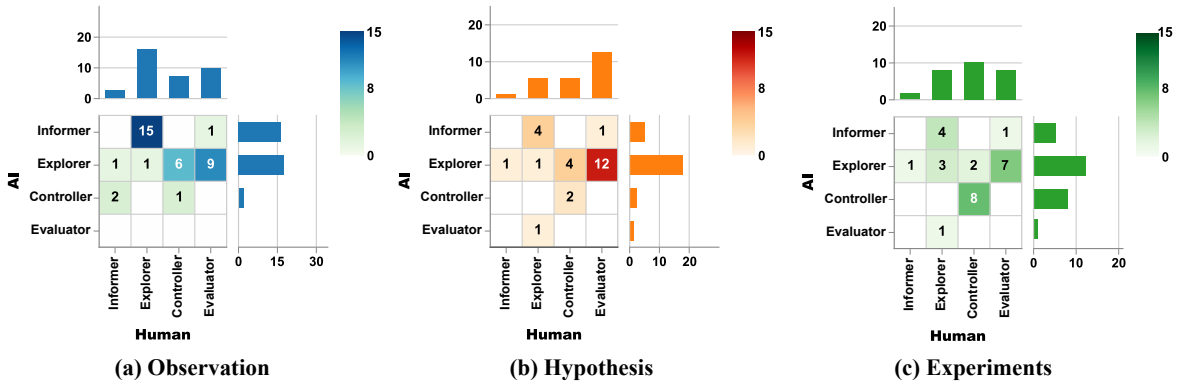


**Figure 2.** The human-AI collaboration patterns across three scientific discovery stages.

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

 doi:10.20944/preprints202601.0405.v2



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 human–AI 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 human–AI 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, this section discusses several significant challenges and potential future research directions for human-AI collaboration in scientific discovery.

### *4.1. 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.

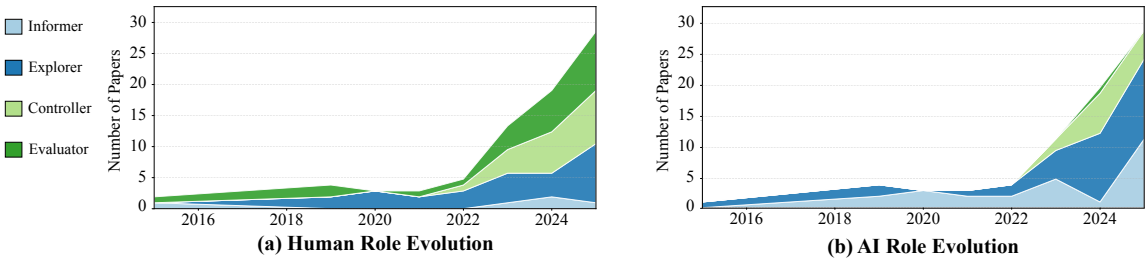


**Figure 3.** Evolution of human and AI roles.

The 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 [67, 68]. (2) Causal and mechanistic reasoning, beyond surface pattern recognition; without understanding the underlying "why", evaluations may be misled by spurious correlations [69]. (3) Contextual and normative judgment, aligning with tacit scientific standards–a challenge reflected in AI's difficulty with complex, value-laden trade-offs [70]. 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

 doi:10.20944/preprints202601.0405.v2



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.

### *4.2. Generative Interfaces for Supporting Human Involvement*

Most existing interfaces for human–AI 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 [71]. 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 human-AI 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 [66,72], 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.

### *4.3. Adaptive Role Assignments Between Human and AI*

Existing human-AI collaborative research typically predefines the roles of human and AI based on their abilities and limitations to solve scientific discovery tasks [25,36]. 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. 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 human–AI 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.

### *4.4. 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 exper-

 doi:10.20944/preprints202601.0405.v2



imental processes. However, in domains like biology, chemistry, and medicine, scientific discovery fundamentally relies on a large number of physical laboratory experiments rather than data analysis alone [73,74]. 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 [7].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.

### *4.5. 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 [75]. 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 human-AI collaboration in scientific discovery, focusing on the roles of humans and AI across the stages of observation, hypothesis, and experiment. By introducing a novel taxonomy of four roles of human-AI 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, including challenges related to role coordination, validation, and transparency. Finally, we outlined a research agenda for developing more adaptive, trustworthy, and efficient human-AI systems.

## 6. Limitations

One limitation of this study is that we only included papers that explicitly address specific problems in the three stages of scientific discovery, while excluding more general-purpose papers, such as those aimed at assisting humans with tasks like writing or data collection. This choice was made because such tools operate outside the three main stages we focus on, serving more preparatory or subsequent roles. Nevertheless, these tools remain important, and future work could explore how they can be integrated into a broader framework of scientific discovery. Another limitation is that our analysis focuses exclusively on the natural sciences. This is because methodologies in the social sciences differ substantially, which may result in human–AI collaboration approaches that do not directly align with those observed in the natural sciences. Future work could explore whether our taxonomy can be adapted or extended to the social sciences.

 doi:10.20944/preprints202601.0405.v2



# Appendix A. Coding Results

| Paper | Stage | | | Human Role | | | | AI Role | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Observation | Hypothesis | Experiment | Informer | Explorer | Evaluator | Controller | Informer | Explorer | Evaluator | Controller |
| A Predictive Visual Analytics System for Studying Neurodegenerative Disease Based on DTI Fiber Tracts | ■ | | | | ■ | | | ■ | | | |
| Bioinformatics Agent (BIA): Unleashing the Power of LLMs to Reshape Bioinformatics Workflow | ■ | | | | | | ■ | | ■ | | |
| CellScout: Visual Analytics for Mining Biomarkers in Cell State Discovery | ■ | | | | ■ | | | ■ | | | ■ |
| ChemVA: Interactive Visual Analysis of Chemical Compound Similarity in Virtual Screening | ■ | | | | ■ | | | ■ | | | |
| ClimateSOM: A Visual Analysis Workflow for Climate Ensemble Datasets | ■ | | | | ■ | | | ■ | | | |
| Completing A Systematic Review in Hours instead of Months with Interactive AI Agents | ■ | | | ■ | | | ■ | ■ | ■ | | |
| DASS Good: Explainable Data Mining of Spatial Cohort Data | ■ | | | | ■ | | ■ | ■ | | | |
| DiffFit: Visually-Guided Differentiable Fitting of Molecule Structures to a Cryo-EM Map | ■ | | | | | ■ | ■ | | ■ | | |
| DTBIA: An Immersive Visual Analytics System for Brain-Inspired Research | ■ | | | | ■ | | | ■ | | | |
| Explaining Air Quality Forecast for Verifying Domain Knowledge using Feature Importance Visualization | ■ | | | | ■ | ■ | | ■ | | | |
| Extending the Nested Model for User-Centric XAI: A Design Study on GNN-based Drug Repurposing | ■ | | | | | ■ | | ■ | ■ | | |
| Facetto: Combining Unsupervised and Supervised Learning for Hierarchical Phenotype Analysis in Multi-Channel Image Data | ■ | | | | ■ | ■ | | ■ | | | |
| GeneAgent: self-verification language agent for gene-set analysis using domain databases | ■ | | | | | ■ | | ■ | ■ | ■ | |
| HealthPrism: A Visual Analytics System for Exploring Children's Physical and Mental Health Profiles with Multimodal Data | ■ | | | | ■ | | | ■ | | | |
| IdMotif: An Interactive Motif Identification in Protein Sequences | ■ | | | | ■ | ■ | | ■ | | | |
| Lessons from the Development of an Anomaly Detection Interface on the Mars Perseverance Rover using the ISHMAP Framework | ■ | | | ■ | | | ■ | ■ | ■ | | |
| Optimal Dimensionality Selection Using Hull Heatmaps for Single-Cell Analysis | ■ | | | | ■ | | | ■ | | | |
| Roses Have Thorns: Understanding the Downside of Oncological Care Delivery Through Visual Analytics and Sequential Rule Mining | ■ | | | | | ■ | ■ | ■ | ■ | | |
| THALIS: Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy | ■ | | | | ■ | ■ | | ■ | ■ | | |
| Visualizing and Comparing Machine Learning Predictions to Improve Human-AI Teaming on the Example of Cell Lineage | ■ | | | ■ | | ■ | | ■ | ■ | | |
| Uncover: Toward Interpretable Models for Detecting New Star Cluster Members | ■ | | | | ■ | | ■ | | ■ | | |
| Cell2Cell: Explorative Cell Interaction Analysis in Multi-Volumetric Tissue Data | ■ | | | | ■ | | | ■ | | | |
| TrajLens: Visual Analysis for Constructing Cell Developmental Trajectories in Cross-Sample Exploration | ■ | | | | ■ | ■ | | | ■ | | |
| Visual Analysis of Multi-Outcome Causal Graphs | ■ | | | | | | ■ | ■ | ■ | | |
| HypoChainer: A Collaborative System Combining LLMs and Knowledge Graphs for Hypothesis-Driven Scientific Discovery | | ■ | | | ■ | | | ■ | ■ | | |
| Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning | | ■ | | ■ | | ■ | | | ■ | | |
| ChatDrug: Conversational Drug Editing with Retrieval and Domain Feedback | | ■ | | | | ■ | | | ■ | | ■ |
| ChemoGraph: Interactive Visual Exploration of the Chemical Space | | ■ | | | ■ | ■ | | | ■ | | |
| DIVA: Exploration and Validation of Hypothesized Drug-Drug Interactions | | ■ | | | | ■ | | ■ | ■ | | |
| DrugAssist: A LLM for Molecule Optimization | | ■ | | | | | ■ | | ■ | | |
| dZiner: Rational Inverse Design of Materials with AI Agents | | ■ | | | | | ■ | | ■ | | ■ |
| matexplorer: Visual exploration on predicting ionic conductivity for solid-state electrolytes | | ■ | | | ■ | ■ | | ■ | | | |
| MedChemLens: An Interactive Visual Tool to Support Direction Selection in Interdisciplinary Experimental Research of Medicinal Chemistry | | ■ | | | ■ | | | ■ | | | |
| Scideator: Human-LLM Scientific Idea Generation Grounded in Research-Paper Facet Recombination | | ■ | | | ■ | ■ | | | ■ | ■ | |
| SLInterpreter: An Exploratory and Iterative Human-AI Collaborative System for GNN-Based Synthetic Lethal Prediction | | ■ | | | | ■ | ■ | ■ | ■ | | |
| SPARK: Harnessing Human-Centered Workflows with Biomedical Foundation Models for Drug Discovery | | ■ | | | | ■ | ■ | | ■ | | |
| Visual Analytics for Hypothesis-Driven Exploration in Computational Pathology | | ■ | | | ■ | | | ■ | | | |
| Human-in-the-loop interface for Automated experiments in Electron Microscopy, Automated characterization | | | ■ | | ■ | | ■ | | ■ | | |
| SimuLearn: Fast and Accurate Simulator to Support Morphing Materials Design and Workflows | | | ■ | | ■ | | | ■ | | | |
| NNVA: Neural Network Assisted Visual Analysis of Yeast Cell Polarization Simulation | | | ■ | | ■ | ■ | | ■ | ■ | | |
| SynthLens: Visual Analytics for Facilitating Multi-Step Synthetic Route Design | | | ■ | | ■ | ■ | | ■ | | | |
| RetroLens: A Human-AI Collaborative System for Multi-step Retrosynthetic Route Planning | | | ■ | | ■ | ■ | | | ■ | | ■ |
| CRISPR-GPT for agentic automation of gene-editing experiments | | | ■ | | | ■ | ■ | | ■ | | ■ |
| ChemCrow: Augmenting LLMs with Chemistry Tools | | | ■ | | | ■ | ■ | | ■ | | ■ |
| ORGANA: A Robotic Assistant for Automated Chemistry Experimentation and Characterization | | | ■ | | | | ■ | | ■ | | ■ |
| An automatic end-to-end chemical synthesis development platform powered by large language models | ■ | ■ | ■ | ■ | | ■ | | ■ | ■ | | ■ |
| SciClaims: An End-to-End Generative System for Biomedical Claim Analysis | ■ | ■ | ■ | | | ■ | | | ■ | ■ | |
| MatAgent: A human-in-the-loop multi-agent LLM framework for accelerating the material science discovery cycle | ■ | ■ | ■ | | | ■ | ■ | | ■ | | ■ |
| The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies | ■ | ■ | ■ | | ■ | ■ | ■ | | ■ | | ■ |
| Towards an AI Co-Scientist | ■ | ■ | ■ | | ■ | ■ | ■ | ■ | ■ | | ■ |
| MatPilot: an LLM-enabled AI Materials Scientist under the Framework of Human-Machine Collaboration | ■ | ■ | ■ | | ■ | ■ | ■ | | ■ | | ■ |

## References


1. Zhang, Y.; Chen, X.; Jin, B.; Wang, S.; Ji, S.; Wang, W.; Han, J. A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 8783–8817. https://doi.org/10.18653/v1/2024.emnlp-main.498.
2. Tang, J.; Xia, L.; Li, Z.; Huang, C. AI-Researcher: Autonomous Scientific Innovation. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
3. Ye, F.; Zheng, Z.; Xue, D.; Shen, Y.; Wang, L.; Ma, Y.; Wang, Y.; Wang, X.; Zhou, X.; Gu, Q. ProteinBench: A Holistic Evaluation of Protein Foundation Models. In Proceedings of the International Conference on Representation Learning; Yue, Y.; Garg, A.; Peng, N.; Sha, F.; Yu, R., Eds., 2025, Vol. 2025, pp. 29857–29891.

 doi:10.20944/preprints202601.0405.v2




4. Zhang, Y.; Wei, Z.; Yuan, Y.; Li, C.; Huang, W. EquiPocket: an E(3)-Equivariant Geometric Graph Neural Network for Ligand Binding Site Prediction. In Proceedings of the Proceedings of the 41st International Conference on Machine Learning; Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; Berkenkamp, F., Eds., 2024, Vol. 235, pp. 60021–60039.
5. Ma, Y.; Gou, Z.; Hao, J.; Xu, R.; Wang, S.; Pan, L.; Yang, Y.; Cao, Y.; Sun, A. SciAgent: Tool-augmented Language Models for Scientific Reasoning. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing; Al-Onaizan, Y.; Bansal, M.; Chen, Y.N., Eds., 2024, pp. 15701–15736. https://doi.org/10.18653/v1/2024.emnlp-main.880.
6. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the International Conference on Learning Representations, 2023.
7. Pratiush, U.; Duscher, G.; Kalinin, S. Human-in-the-loop interface for Automated experiments in Electron Microscopy, Automated characterization. In Proceedings of the AI for Accelerated Materials Design - NeurIPS 2024, 2024.
8. Jansen, P.; Tafjord, O.; Radensky, M.; Siangliulue, P.; Hope, T.; Dalvi Mishra, B.; Majumder, B.P.; Weld, D.S.; Clark, P. CodeScientist: End-to-End Semi-Automated Scientific Discovery with Code-based Experimentation. 2025, pp. 13370–13467. https://doi.org/10.18653/v1/2025.findings-acl.692.
9. Zheng, T.; Deng, Z.; Tsang, H.T.; Wang, W.; Bai, J.; Wang, Z.; Song, Y. From Automation to Autonomy: A Survey on Large Language Models in Scientific Discovery. 2025, pp. 17744–17761. https://doi.org/10.18653/v1/2025.emnlp-main.895.
10. Reddy, C.K.; Shojaee, P. Towards scientific discovery with generative AI: progress, opportunities, and challenges. In Proceedings of the Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence. AAAI Press, 2025. https://doi.org/10.1609/aaai.v39i27.35084.
11. Huang, C.; Deng, Y.; Lei, W.; Lv, J.; Chua, T.S.; Huang, J. How to Enable Effective Cooperation Between Humans and NLP Models: A Survey of Principles, Formalizations, and Beyond. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., 2025, pp. 466–488. https://doi.org/10.18653/v1/2025.acl-long.22.
12. Rajashekar, N.C.; Shin, Y.E.; Pu, Y.; Chung, S.; You, K.; Giuffre, M.; Chan, C.E.; Saarinen, T.; Hsiao, A.; Sekhon, J.; et al. Human-Algorithmic Interaction Using a Large Language Model-Augmented Artificial Intelligence Clinical Decision Support System. In Proceedings of the Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024, CHI '24. https://doi.org/10.1145/3613904.3642024.
13. Li, H.; Wang, Y.; Qu, H. Where Are We So Far? Understanding Data Storytelling Tools from the Perspective of Human-AI Collaboration. In Proceedings of the Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024. https://doi.org/10.1145/3613904.3642726.
14. Mohanty, V.; Lim, J.; Luther, K. What Lies Beneath? Exploring the Impact of Underlying AI Model Updates in AI-Infused Systems. In Proceedings of the Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, 2025, CHI '25. https://doi.org/10.1145/3706598.3713751.
15. Wang, H.; Fu, T.; Du, Y.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Van Katwyk, P.; Deac, A.; et al. Scientific discovery in the age of artificial intelligence. *Nature* **2023**, *620*, 47–60. https://doi.org/10.1038/s41586-023-06221-2.
16. Wang, X.; Huey, S.L.; Sheng, R.; Mehta, S.; Wang, F. SciDaSynth: Interactive Structured Data Extraction From Scientific Literature With Large Language Model. *Campbell Systematic Reviews* **2025**, *21*, e70073. https://doi.org/10.1002/cl2.70073.
17. Floricel, C.; Nipu, N.; Biggs, M.; Wentzel, A.; Canahuate, G.; Van Dijk, L.; Mohamed, A.; Fuller, C.; Marai, G. THALIS: Human-Machine Analysis of Longitudinal Symptoms in Cancer Therapy. *IEEE Transactions on Visualization and Computer Graphics* **2022**, *28*, 151–161. https://doi.org/10.1109/TVCG.2021.3114810.
18. Wright, A.P.; Nemere, P.; Galvin, A.; Chau, D.H.; Davidoff, S. Lessons from the Development of an Anomaly Detection Interface on the Mars Perseverance Rover using the ISHMAP Framework. In Proceedings of the Proceedings of the 28th International Conference on Intelligent User Interfaces, 2023, p. 91–105. https://doi.org/10.1145/3581641.3584036.
19. Sabando, M.V.; Ulbrich, P.; Selzer, M.; Byška, J.; Mičan, J.; Ponzoni, I.; Soto, A.J.; Ganuza, M.L.; Kozlíková, B. ChemVA: Interactive Visual Analysis of Chemical Compound Similarity in Virtual Screening. *IEEE Transactions on Visualization and Computer Graphics* **2021**, *27*, 891–901. https://doi.org/10.1109/TVCG.2020.3030438.

 doi:10.20944/preprints202601.0405.v2




20. Shi, C.; Hu, Y.; Wang, S.; Ma, S.; Zheng, C.; Ma, X.; Luo, Q. RetroLens: A Human-AI Collaborative System for Multi-step Retrosynthetic Route Planning. 2023. https://doi.org/10.1145/3544548.3581469.
21. Xin, Q.; Kong, Q.; Ji, H.; Shen, Y.; Liu, Y.; Sun, Y.; Zhang, Z.; Li, Z.; Xia, X.; Deng, B.; et al. BioInformatics Agent (BIA): Unleashing the Power of Large Language Models to Reshape Bioinformatics Workflow. *bioRxiv* **2024**, [https://www.biorxiv.org/content/early/2024/05/22/2024.05.22.595240.full.pdf]. https://doi.org/10.1101/2024.05.22.595240.
22. Kawakami, Y.; Cayan, D.; Liu, D.; Ma, K.L. ClimateSOM: a Visual Analysis Workflow for Climate Ensemble Datasets. *IEEE Transactions on Visualization and Computer Graphics* **2025**, pp. 1–11. https://doi.org/10.1109/TVCG.2025.3634788.
23. Jeong, H.; Jeong, H.o.; Lee, S.; Jeong, W.K. Optimal Dimensionality Selection Using Hull Heatmaps for Single-Cell Analysis. *Computer Graphics Forum* **2025**, *44*, e70151. https://doi.org/10.1111/cgf.70151.
24. Mörth, E.; Sidak, K.; Maliga, Z.; Möller, T.; Gehlenborg, N.; Sorger, P.; Pfister, H.; Beyer, J.; Krüger, R. Cell2Cell: Explorative Cell Interaction Analysis in Multi-Volumetric Tissue Data. *IEEE Transactions on Visualization and Computer Graphics* **2025**, *31*, 569–579. https://doi.org/10.1109/TVCG.2024.3456406.
25. Jiang, Z.; Chen, H.; Zhou, R.; Deng, J.; Zhang, X.; Zhao, R.; Xie, C.; Wang, Y.; Ngai, E.C. HealthPrism: A Visual Analytics System for Exploring Children's Physical and Mental Health Profiles with Multimodal Data . *IEEE Transactions on Visualization & Computer Graphics* **2024**, *30*, 1205–1215. https://doi.org/10.1109/TVCG.2023.3326943.
26. Yao, J.H.; Li, M.; Liu, J.; Li, Y.; Feng, J.; Han, J.; Zheng, Q.; Feng, J.; Chen, S. DTBIA: An Immersive Visual Analytics System for Brain-Inspired Research. *IEEE Transactions on Visualization and Computer Graphics* **2025**, *31*, 3796–3808. https://doi.org/10.1109/TVCG.2025.3567135.
27. Wang, Q.; Ruan, S.; Sheng, R.; Wang, Y.; Zhu, M.; Qu, H. TrajLens: Visual Analysis for Constructing Cell Developmental Trajectories in Cross-Sample Exploration. *IEEE Transactions on Visualization and Computer Graphics* **2025**, pp. 1–11. https://doi.org/10.1109/TVCG.2025.3634875.
28. Sheng, R.; Zang, Z.; Wang, J.; Luo, Y.; Chen, Z.; Zhou, Y.; Ruan, S.; Qu, H. CellScout: Visual Analytics for Mining Biomarkers in Cell State Discovery. *IEEE Transactions on Visualization and Computer Graphics* **2025**, pp. 1–16. https://doi.org/10.1109/TVCG.2025.3636102.
29. Park, J.H.; Prasad, V.; Newsom, S.; Najar, F.; Rajan, R. IdMotif: An Interactive Motif Identification in Protein Sequences. *IEEE Computer Graphics and Applications* **2024**, *44*, 114–125. https://doi.org/10.1109/MCG.2023.3345742.
30. Palaniyappan Velumani, R.; Xia, M.; Han, J.; Wang, C.; LAU, A.K.; Qu, H. AQX: Explaining Air Quality Forecast for Verifying Domain Knowledge using Feature Importance Visualization. In Proceedings of the Proceedings of the 27th International Conference on Intelligent User Interfaces, 2022, p. 720–733. https://doi.org/10.1145/3490099.3511150.
31. Xu, C.; Neuroth, T.; Fujiwara, T.; Liang, R.; Ma, K.L. A Predictive Visual Analytics System for Studying Neurodegenerative Disease Based on DTI Fiber Tracts. *IEEE Transactions on Visualization and Computer Graphics* **2023**, *29*, 2020–2035. https://doi.org/10.1109/TVCG.2021.3137174.
32. Krueger, R.; Beyer, J.; Jang, W.D.; Kim, N.W.; Sokolov, A.; Sorger, P.K.; Pfister, H. Facetto: Combining Unsupervised and Supervised Learning for Hierarchical Phenotype Analysis in Multi-Channel Image Data. *IEEE Transactions on Visualization and Computer Graphics* **2020**, *26*, 227–237. https://doi.org/10.1109/TVCG.2019.2934547.
33. Wang, Q.; Huang, K.; Chandak, P.; Zitnik, M.; Gehlenborg, N. Extending the Nested Model for User-Centric XAI: A Design Study on GNN-based Drug Repurposing. *IEEE Transactions on Visualization and Computer Graphics* **2023**, *29*, 1266–1276. https://doi.org/10.1109/TVCG.2022.3209435.
34. Luo, D.; Alsuwaykit, Z.; Khan, D.; Strnad, O.; Isenberg, T.; Viola, I. DiffFit: Visually-Guided Differentiable Fitting of Molecule Structures to a Cryo-EM Map. *IEEE Transactions on Visualization and Computer Graphics* **2025**, *31*, 558–568. https://doi.org/10.1109/TVCG.2024.3456404.
35. Floricel, C.; Wentzel, A.; Mohamed, A.; Fuller, C.; Canahuate, G.; Marai, G. Roses Have Thorns: Understanding the Downside of Oncological Care Delivery Through Visual Analytics and Sequential Rule Mining. *IEEE Transactions on Visualization and Computer Graphics* **2024**, *30*, 1227–1237. https://doi.org/10.1109/TVCG.2023.3326939.
36. Hong, J.; Maciejewski, R.; Trubuil, A.; Isenberg, T. Visualizing and Comparing Machine Learning Predictions to Improve Human-AI Teaming on the Example of Cell Lineage. *IEEE Transactions on Visualization and Computer Graphics* **2024**, *30*, 1956–1969. https://doi.org/10.1109/TVCG.2023.3302308.

 doi:10.20944/preprints202601.0405.v2




37. Wang, Z.; Jin, Q.; Wei, C.H.; Tian, S.; Lai, P.T.; Zhu, Q.; Day, C.P.; Ross, C.; Leaman, R.; Lu, Z. GeneAgent: Self-Verification Language Agent for Gene-Set Analysis Using Domain Databases. *Nature Methods* **2025**, *22*, 1677–1685. https://doi.org/10.1038/s41592-025-02748-6.
38. Qiu, R.; Chen, S.; Su, Y.; Yen, P.Y.; Shen, H.W. Completing A Systematic Review in Hours instead of Months with Interactive AI Agents. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Che, W.; Nabende, J.; Shutova, E.; Pilehvar, M.T., Eds., 2025, pp. 31559–31593. https://doi.org/10.18653/v1/2025.acl-long.1523.
39. Ratzenböck, S.; Obermüller, V.; Möller, T.; Alves, J.; Bomze, I.M. Uncover: Toward Interpretable Models for Detecting New Star Cluster Members. *IEEE Transactions on Visualization and Computer Graphics* **2023**, *29*, 3855–3872. https://doi.org/10.1109/TVCG.2022.3172560.
40. Wentzel, A.; Floricel, C.; Canahuate, G.; Naser, M.A.; Mohamed, A.S.; Fuller, C.D.; van Dijk, L.; Marai, G.E. DASS Good: Explainable Data Mining of Spatial Cohort Data. *Computer Graphics Forum* **2023**, *42*, 283–295. https://doi.org/10.1111/cgf.14830.
41. Fan, M.; Yu, J.; Weiskopf, D.; Cao, N.; Wang, H.Y.; Zhou, L. Visual Analysis of Multi-Outcome Causal Graphs. *IEEE Transactions on Visualization and Computer Graphics* **2025**, *31*, 656–666. https://doi.org/10.1109/TVCG.2024.3456346.
42. Liu, S.; Wang, J.; Yang, Y.; Wang, C.; Liu, L.; Guo, H.; Xiao, C. Conversational Drug Editing Using Retrieval and Domain Feedback. In Proceedings of the The Twelfth International Conference on Learning Representations (ICLR), 2024.
43. Lu, S.; Zhou, Q.; Ouyang, Y.; Guo, Y.; Li, Q.; Wang, J. Accelerated Discovery of Stable Lead-Free Hybrid Organic-Inorganic Perovskites via Machine Learning. *Nature Communications* **2018**, *9*, 3405. https://doi.org/10.1038/s41467-018-05761-w.
44. Ni, Z.; Li, Y.; Hu, K.; Han, K.; Xu, M.; Chen, X.; Liu, F.; Ye, Y.; Bai, S. MatPilot: An LLM-Enabled AI Materials Scientist Under the Framework of Human-Machine Collaboration. *arXiv preprint arXiv:2411.08063* **2024**.
45. Kale, B.; Clyde, A.; Sun, M.; Ramanathan, A.; Stevens, R.; Papka, M.E. ChemoGraph: Interactive Visual Exploration of the Chemical Space. 2023, Vol. 42, pp. 13–24. https://doi.org/10.1111/cgf.14807.
46. Swanson, K.; Wu, W.; Bulaong, N.L.; Pak, J.E.; Zou, J. The Virtual Lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature* **2025**, *646*, 716–723. https://doi.org/10.1038/s41586-025-09442-9.
47. Wang, Q.; Schaeffer, R.; Khani, F.; Ilievski, F.; et al. Scideator: Human-LLM Scientific Idea Generation Grounded in Research-Paper Facet Recombination. *Transactions on Machine Learning Research* **2025**.
48. Kakar, T.; Qin, X.; Rundensteiner, E.A.; Harrison, L.; Sahoo, S.K.; De, S. DIVA: Exploration and Validation of Hypothesized Drug-Drug Interactions. In Proceedings of the Computer Graphics Forum, 2019, Vol. 38, pp. 95–106. https://doi.org/10.1111/cgf.13676.
49. Ortega, R.; Gomez-Perez, J.M. SciClaims: An End-to-End Generative System for Biomedical Claim Analysis. In Proceedings of the Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations; Habernal, I.; Schulam, P.; Tiedemann, J., Eds., 2025, pp. 141–154. https://doi.org/10.18653/v1/2025.emnlp-demos.11.
50. Ansari, M.; Watchorn, J.; Brown, C.E.; Brown, J.S. dZiner: Rational Inverse Design of Materials with AI Agents. In Proceedings of the AI for Accelerated Materials Design - NeurIPS 2024, 2024.
51. Ye, G.; Cai, X.; Lai, H.; Wang, X.; Huang, J.; Wang, L.; Liu, W.; Zeng, X. DrugAssist: a large language model for molecule optimization. *Briefings in Bioinformatics* **2025**, *26*, bbae693. https://doi.org/10.1093/bib/bbae693.
52. Kwon, B.C.; Rabinovici-Cohen, S.; Moturi, B.; Mwaura, R.; Wahome, K.; Njeru, O.; Shinyenyi, M.; Wanjiru, C.; Remy, S.; Ogallo, W.; et al. SPARK: harnessing human-centered workflows with biomedical foundation models for drug discovery. 2024, IJCAI '24. https://doi.org/10.24963/ijcai.2024/1015.
53. Jiang, H.; Shi, S.; Zhang, S.; Zheng, J.; Li, Q. SLInterpreter: An Exploratory and Iterative Human-AI Collaborative System for GNN-Based Synthetic Lethal Prediction. *IEEE Transactions on Visualization and Computer Graphics* **2025**, *31*, 919–929. https://doi.org/10.1109/TVCG.2024.3456325.
54. Jiang, H.; Shi, S.; Yao, Y.; Jiang, C.; Li, Q. HypoChainer: a Collaborative System Combining LLMs and Knowledge Graphs for Hypothesis-Driven Scientific Discovery. *IEEE Transactions on Visualization and Computer Graphics* **2025**, pp. 1–11. https://doi.org/10.1109/TVCG.2025.3633887.
55. Shi, C.; Nie, F.; Hu, Y.; Xu, Y.; Chen, L.; Ma, X.; Luo, Q. MedChemLens: An Interactive Visual Tool to Support Direction Selection in Interdisciplinary Experimental Research of Medicinal Chemistry. *IEEE Transactions on Visualization and Computer Graphics* **2023**, *29*, 63–73. https://doi.org/10.1109/TVCG.2022.3209434.

 doi:10.20944/preprints202601.0405.v2




56. Corvò, A.; Caballero, H.S.G.; Westenberg, M.A.; van Driel, M.A.; van Wijk, J.J. Visual Analytics for Hypothesis-Driven Exploration in Computational Pathology. *IEEE Transactions on Visualization and Computer Graphics* **2021**, *27*, 3851–3866. https://doi.org/10.1109/TVCG.2020.2990336.
57. Pu, J.; Shao, H.; Gao, B.; Zhu, Z.; Zhu, Y.; Rao, Y.; Xiang, Y. matExplorer: Visual Exploration on Predicting Ionic Conductivity for Solid-state Electrolytes. *IEEE Transactions on Visualization and Computer Graphics* **2022**, *28*, 65–75. https://doi.org/10.1109/TVCG.2021.3114812.
58. Darvish, K.; Skreta, M.; Zhao, Y.; Yoshikawa, N.; Som, S.; Bogdanovic, M.; Cao, Y.; Hao, H.; Xu, H.; Aspuru-Guzik, A.; et al. ORGANA: A robotic assistant for automated chemistry experimentation and characterization. *Matter* **2025**, *8*, 101897. https://doi.org/https://doi.org/10.1016/j.matt.2024.10.015.
59. Bran, A.M.; Cox, S.; Schilter, O.; Baldassari, C.; White, A.D.; Schwaller, P. Augmenting large language models with chemistry tools. *Nature Machine Intelligence* **2024**, *6*, 525–535. https://doi.org/10.1038/s42256-024-00832-8.
60. Boiko, D.A.; MacKnight, R.; Kline, B.; Gomes, G. Autonomous chemical research with large language models. *Nature* **2023**, *624*, 570–578. https://doi.org/10.1038/s41586-023-06792-0.
61. Qu, Y.; Huang, K.; Yin, M.; Zhan, K.; Liu, D.; Yin, D.; Cousins, H.C.; Johnson, W.A.; Wang, X.; Shah, M.; et al. CRISPR-GPT for agentic automation of gene-editing experiments. *Nature Biomedical Engineering* **2025**. https://doi.org/10.1038/s41551-025-01463-z.
62. Bazgir, A.; chandra Praneeth Madugula, R.; Zhang, Y. MatAgent: A human-in-the-loop multi-agent LLM framework for accelerating the material science discovery cycle. In Proceedings of the AI for Accelerated Materials Design - ICLR 2025, 2025.
63. Gottweis, J.; Weng, W.H.; Daryin, A.; Tu, T.; Palepu, A.; Sirkovic, P.; Myaskovsky, A.; Weissenberger, F.; Rong, K.; Tanno, R.; et al. Towards an AI Co-Scientist. *arXiv preprint arXiv:2502.18864* **2025**.
64. Yang, H.; Qian, K.; Liu, H.; Yu, Y.; Gu, J.; McGehee, M.; Zhang, Y.J.; Yao, L. SimuLearn: Fast and Accurate Simulator to Support Morphing Materials Design and Workflows. In Proceedings of the Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology, 2020, UIST '20, p. 71–84. https://doi.org/10.1145/3379337.3415867.
65. Hazarika, S.; Li, H.; Wang, K.C.; Shen, H.W.; Chou, C.S. NNVA: Neural Network Assisted Visual Analysis of Yeast Cell Polarization Simulation. *IEEE Transactions on Visualization and Computer Graphics* **2020**, *26*, 34–44. https://doi.org/10.1109/TVCG.2019.2934591.
66. Wang, Q.; Sheng, R.; Ruan, S.; Jin, X.; Shi, C.; Zhu, M. SynthLens: Visual Analytics for Facilitating Multi-Step Synthetic Route Design. *IEEE Transactions on Visualization and Computer Graphics* **2025**, *31*, 7647–7660. https://doi.org/10.1109/TVCG.2025.3552134.
67. Xie, Q.; Li, Q.; Yu, Z.; Zhang, Y.; Zhang, Y.; Yang, L. An Empirical Analysis of Uncertainty in Large Language Model Evaluations. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
68. Heo, J.; Xiong, M.; Heinze-Deml, C.; Narain, J. Do LLMs estimate uncertainty well in instruction-following? In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
69. Chen, Y.; Zhang, C.; Luo, D.; D'Haro, L.F.; Tan, R.; Li, H. Unveiling the Achilles' Heel of NLG Evaluators: A Unified Adversarial Framework Driven by Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 1359–1375. https://doi.org/10.18653/v1/2024.findings-acl.80.
70. Rezaei, M.; Fu, Y.; Cuvin, P.; Ziems, C.; Zhang, Y.; Zhu, H.; Yang, D. EgoNormia: Benchmarking Physical-Social Norm Understanding. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 19256–19283. https://doi.org/10.18653/v1/2025.findings-acl.985.
71. Chen, J.; Zhang, Y.; Zhang, Y.; Shao, Y.; Yang, D. Generative Interfaces for Language Models. *arXiv preprint arXiv:2508.19227* **2025**.
72. Sheng, R.; Wang, X.; Wang, J.; Jin, X.; Sheng, Z.; Xu, Z.; Rajendran, S.; Qu, H.; Wang, F. TrialCompass: Visual Analytics for Enhancing the Eligibility Criteria Design of Clinical Trials. *IEEE Transactions on Visualization and Computer Graphics* **2025**, pp. 1–11. https://doi.org/10.1109/TVCG.2025.3634803.
73. Luro, S.; Potvin-Trottier, L.; Okumus, B.; Paulsson, J. Isolating live cells after high-throughput, long-term time-lapse microscopy. *Nature Methods* **2020**, *17*, 93–100. https://doi.org/10.1038/s41592-019-0642-5.
74. Wright, P.M.; Seiple, I.B.; Myers, A.G. The Evolving Role of Chemical Synthesis in Antibacterial Drug Discovery. *Angewandte Chemie International Edition* **2014**, *53*, 8840–8869. https://doi.org/10.1002/anie.201310843.

 doi:10.20944/preprints202601.0405.v2



75. Schmidgall, S.; Su, Y.; Wang, Z.; Sun, X.; Wu, J.; Yu, X.; Liu, J.; Moor, M.; Liu, Z.; Barsoum, E. Agent Laboratory: Using LLM Agents as Research Assistants. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2025; Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; Peng, V., Eds., 2025, pp. 5977–6043. https://doi.org/10.18653/v1/2025.findings-emnlp.320.