








A conceptual framework for human–AI collaborative genome annotation

Xiaomei Li ^{1,*}, Alex Whan ², Meredith McNeil ³, David Starns ⁴, Jessica Irons ⁵, Samuel C. Andrew ¹, Rad Suchecki ^{6,7,*}

¹Agriculture and Food, CSIRO, 26 Pembroke Road, Marsfield, NSW 2122, Australia

²Agriculture and Food, CSIRO, 2–40 Clunies Ross Street, Acton, ACT 2601, Australia

³Agriculture and Food, CSIRO, 306 Carmody Road, St Lucia, QLD 4067, Australia

⁴School of Molecular and Cellular Biology, University of Leeds, Woodhouse Lane, Leeds LS2 9JT, United Kingdom

⁵Data 61, CSIRO, 13 Garden Street, Eveleigh, NSW 2015, Australia

⁶Agriculture and Food, CSIRO, 40 Waite Road, Urrbrae, SA 5064, Australia

⁷Present address: Alkahest Inc., 125 Shoreway Rd D, San Carlos, CA 94070, United States

*Corresponding authors. Xiaomei Li, Agriculture and Food, CSIRO, 26 Pembroke Road, Marsfield, NSW 2122, Australia. E-mail: Maisie.Li@csiro.au; Rad Suchecki, E-mail: rsuchecki@alkahest.com

Abstract

Genome annotation is essential for understanding the functional elements within genomes. While automated methods are indispensable for processing large-scale genomic data, they often face challenges in accurately predicting gene structures and functions. Consequently, manual curation by domain experts remains crucial for validating and refining these predictions. These combined outcomes from automated tools and manual curation highlight the importance of integrating human expertise with artificial intelligence (AI) capabilities to improve both the accuracy and efficiency of genome annotation. However, the manual curation process is inherently labor-intensive and time-consuming, making it difficult to scale for large datasets. To address these challenges, we propose a conceptual framework, Human-AI Collaborative Genome Annotation (HAICoGA), that leverages the synergistic partnership between humans and AI to enhance human capabilities and accelerate the genome annotation process. Additionally, we explore the potential of integrating large language models into this framework to support and augment specific tasks. Finally, we discuss emerging challenges and outline open research questions to guide further exploration in this area.

Keywords: genome annotation; human; artificial intelligence; collaboration; conceptual framework; large language model

Introduction

Genome annotation (GA) is the process of identifying and interpreting the functional elements encoded within a genome. It is a critical step in understanding an organism's biology, enabling researchers to connect genetic information to phenotypes, understand disease mechanisms, and uncover evolutionary relationships. GA heavily relies on automated methods, including machine learning (ML) [1–3] and other computational methods such as rule-based and heuristic methods [4, 5]. However, automated methods are generally hampered by the relative scarcity of reliable labeled data and the complexity of biological systems. In fact, gene annotations, particularly functional annotations, are mostly transferred from one species to another in an automated manner, relying mainly on the similarity of underlying nucleotide sequences, or the corresponding protein sequences.

Manual curation is widely recognized as essential for improving the reliability and accuracy of GA [6–8]. It involves human experts reviewing and refining annotations, particularly by addressing ambiguities or gaps that automated pipelines may overlook. For instance, curators enhance the functional

annotations of genes by incorporating new insights from scientific literature that detail experimental results related to gene function. Additionally, manual curation enables precise gene structure annotation by reviewing evidence from multiple sources, such as omics datasets and experimental outcomes, to accurately define gene boundaries [9].

Despite their value, these evidence sources are often scattered across multiple platforms or embedded within vast datasets, making manual curation a time-consuming and labor-intensive process. Consequently, current GA practices rely heavily on automated annotation, which is not always followed by manual curation [10].

Manual curation has mostly been conducted in cases where teams of annotators collaborate to create accurate and up-to-date annotations for high-priority species or specific gene sets. Correspondingly, computational tools and platforms have been developed to support collaboration among annotators. These tools include algorithms that identify problematic annotations and prioritize them for review [11], as well as platforms that facilitate seamless communication, data sharing, and coordination among annotators, regardless of their geographic location [12, 13].

Received: March 19, 2025. Revised: June 18, 2025. Accepted: July 6, 2025

© The Author(s) 2025. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

However, these tools operate independently and have not addressed the issue of dispersed data sources across various platforms. Additionally, there is a lack of dynamic interaction between the tools and users, i.e. the tools typically run automatically without a user in the loop.

As highlighted by Mac et al. [14], the disconnect between artificial intelligence (AI) tools and their users can potentially impact the effective utilization of AI. They suggest integrating scientists into the loop and combining their expertise with interactive ML to accelerate scientific progress. There is a growing body of research on how humans and AI collaborate to drive advancements in fields such as medicine [15–17] and chemistry [18], particularly through the adoption of large language model (LLMs) (see LLM in the Glossary section). Furthermore, the emerging concept of human–AI collaborative intelligence focus on the combination of humans and AIs working together to solve problems, leveraging the strengths of both parties and enhancing each other’s capabilities [19, 20]. Although still in its early stages, an increasing number of studies demonstrate that human–AI collaboration (HAIC) can lead to superior performance in accomplishing complex tasks [19, 21, 22].

Inspired by these work, we propose a conceptual framework named Human-AI Collaborative Genome Annotation (HAICoGA), in which humans and AI systems not only work interdependently but also collaborate over a sustained period. Collaboration in this context refers to effective functional integration between humans and AI systems. AI systems generate annotation suggestions by leveraging automated GA tools and relevant resources, while human experts review and refine these suggestions to ensure alignment with biological context and domain knowledge. This process is not a one-time interaction but an iterative collaboration, in which humans and AI systems continuously inform each other, enhancing both the accuracy and usability of AI support tools. Facilitating this collaboration is the use of LLM-based agentic systems, which integrate multiple tools and resources into a unified, interactive platform, streamlining the GA workflow and reducing the burden of tool switching or manual integration.

The remainder of this paper is organized as follows. Section 2 provides the necessary background and reviews related work relevant to this study. Section 3 introduces the conceptual framework of HAICoGA, identifying key components and critical capabilities required to establish an effective and sustainable human–AI collaborative relationship. In Section 4, we explore current applications of LLM-based AI agents in the biological and biomedical domains and present a vision for the HAICoGA workflow. Section 5 outlines key future research directions to further realize HAICoGA. We hope this work contributes to the development of human–AI collaborative workflows for GA in the future.

Background and related work

Genome annotation

GA can be interpreted as multidimensional, spanning from the nucleotide level to the biological system level [23]. Genomic elements of interest include, but are not limited to, single nucleotide polymorphisms, coding genes, noncoding genes, regulatory elements, and other noncoding regions. Structural annotation primarily focuses on delineating the physical regions of genomic elements. While the structural annotation offers initial clues, a definitive understanding of functions still requires in-depth analysis.

GA encompasses a broad range of tasks that are now primarily accomplished through various computational approaches

utilizing diverse data types. We provide a rough chronology of the emergence and prominence of different automated methods in [Supplementary Note 1](#). These automated methods can be integrated into highly complex pipelines to perform multiple steps in automated GA. Although automated approaches dominate GA, they still face serious limitations and challenges ([Supplementary Note 1](#)).

Due to the limitations of current computational tools, automated GA frequently produces erroneous results. In particular, genes from non-model organisms are often assigned functions based on homologs or labeled with vague terms such as “hypothetical gene” or “expressed protein,” offering little insight into their biological roles. These inaccuracies not only affect immediate interpretations but also propagate through downstream analyses [24], where they can be further amplified by ML or AI models trained on these data [25]. This creates a feedback loop in which low-quality annotations degrade the reliability of both current databases and future research that depends on them.

Manual curation

Manual curation has primarily been done in model species to continuously improve the accuracy and coverage of their GAs. For example, projects such as HAVANA for the human genome, TIGR for *Arabidopsis thaliana*, and ITAG for *Solanum lycopersicum* produce high-quality annotations manually curated by specialized experts. Manual curation is not limited to collaborative efforts or decentralized networks (detailed manual curation models are provided in [Supplementary Note 2](#)); it also plays a crucial role in individual research. Researchers may engage in manual curation before formulating hypotheses for their studies or when interpreting their results.

Manual curation is an ongoing process that requires the continuous repetition of five general steps [26, 27] (see [Supplementary Note 2](#)). This process is time- and labor-intensive, but it can be made more efficient with the assistance of software tools.

For example, Apollo has been widely used in the GA community and continues to be actively updated [12, 28, 29]. Its current web-based interface supports real-time collaborative annotation and integrates JBrowse [30] for fast, scalable genome visualization. Within Apollo, users can edit gene models, adding or deleting exons, adjusting boundaries, and assigning annotations. But functional annotations still rely on external tools, such as text mining systems. Tools like PubTator Center [31] assist in extracting biological entities and gene functions from literature, yet manual curation remains challenging. This is evident in efforts like BioCreative IV [32], which highlight the expertise and validation required to curate functions even with automated support [33].

More details on additional tools can be found in [Supplementary Note 2](#). These tools are distributed across different platforms. Additionally, automated GA tools operate independently from these manual curation tools. As a result, humans need to spend a significant amount of time running and navigating multiple tools, as well as transferring data between them. To accelerate the GA process, an integrated system is needed, which connects all necessary automated and manual GA tools and enables seamless collaboration between humans and AI tools.

Human–AI collaboration

AI systems can play different roles in human–AI teaming, including automation, augmentation, and collaboration [34]. In automation, AI independently performs tasks without human intervention. In augmentation, AI enhances human experts’ abilities in their tasks. Collaboration refers to humans and AI

working together in a coordinated effort toward overall goals, enabling better outcomes than either could achieve alone. Similarly, Gao *et al.* [15] classify AI systems into four intelligence levels based on their capabilities in hypothesis generation, experimentation, and reasoning. Level 0 consists of AI models used by humans as automated tools. Level 1 includes AI assistants that execute tasks specified by scientists. Level 2 consists of AI collaborators that work alongside scientists to refine hypotheses and utilize a broader array of tools for experimentation and discovery. Level 3 represents AI scientists, which exhibit the highest level of intelligence.

In the current GA workflow, automated GA tools can be categorized as Level 0 AI models. Recently, LLMs have been used to develop various AI assistants (Level 1) in the biological and biomedical domains, which could potentially be adapted for GA. However, a significant gap remains due to the lack of HAIC in GA (Level 2).

The emergence of HAIC frameworks and taxonomies across various domains has advanced our understanding of effective HAIC. For instance, Hartikainen *et al.* [35] propose a framework for designing collaborative systems in smart manufacturing, while Martin *et al.* [36] present a model for human–AI co-design in Design Space Exploration. Dubey *et al.* [37] introduce a framework for successful human–AI teaming in contact centers. Despite this progress, applying HAIC to GA presents unique challenges. We build on these frameworks to introduce HAICoGA and propose future research directions for collaborative GA.

A conceptual framework of human–AI collaborative GA

To support collaborative GA, HAICoGA incorporates and extends key concepts from general HAIC theories [35, 37, 38]. As shown in Fig. 1, it includes seven key elements: humans, AI systems and tools, data, goals and tasks, the human–machine interface, environment, and collaboration. Humans and AI operate as a collaborative team within a dynamic environment. Tasks and goals are continuously updated as they interact via a shared interface, leveraging available data to achieve genome annotation outcomes.

Key elements

Humans

Humans refers to individuals involved in GA, particularly those engaged in manual curation. This group may include biological researchers, experimental scientists, biocurators, and trained students. These individuals bring diverse backgrounds, domain knowledge, and methodological experience to the decision-making process [39]. Research suggests experts rely heavily on their previous experience on the task, as well as their deep and often tacit knowledge of the domain [40]. Decision-makers often employ mental shortcuts and heuristics that are efficient but can be prone to cognitive bias [41, 42].

Collaborating with LLMs has the potential to mitigate individual biases when the complementary strengths of human judgment and machine predictions are effectively leveraged [43]. Still, bias in human curation remains an open and ongoing challenge in the biological domain. For example, the BC4GO study [44] demonstrated substantial variability in manual GO annotation, with low inter-annotator agreement even among trained curators, highlighting that human bias persists regardless of AI involvement and must be continually recognized and mitigated.

While humans may not match AI in processing speed or handling vast datasets, they possess the ability to interpret nuanced

context, adapt to shifting task goals, and act flexibly in dynamic environments [40]. This adaptability is crucial in GA, where the context and requirements may change based on new findings or experimental results. Humans can also perform sanity checks and intervene when AI-generated annotations deviate from established rules or expectations, helping to limit the propagation of errors throughout the system [45]. Consequently, while AI can assist and augment decision-making, the role of experienced human curators remains indispensable [46].

AI systems and tools

The AI element consists of a collection of systems and tools designed to perform or assist in GA. By using the term “AI,” we encompass a broad range of computational tools that can be integrated into intelligent systems to enhance collaboration with humans during GA tasks. Ideally, the AI component should consist of an ecosystem of tools with diverse designs and functionalities, each contributing uniquely to the annotation process.

Automation-focused AI tools streamline the multi-step process of identifying and classifying genes and other functional genomic elements, as seen in established GA pipelines and automation methods [5, 47]. In contrast, augmentation tools are designed to assist human annotators by enhancing their efficiency and effectiveness, for example, Apollo [12]. Collaboration tools could support bidirectional interaction between humans and AI systems.

While certain biases inherent in AI systems (discussed in [Supplementary Note 1](#)) cannot be eliminated, they can be mitigated through “the wisdom of the crowd” approach that combines the complementary strengths of both humans and diverse AI systems.

Data

HAICoGA involves two main types of data: those used for GA and those that support HAIC. Genomic sequence data play an important role in GA, including genome assemblies, expressed sequence tags (ESTs), complementary DNAs (cDNAs), RNA-seq data, and protein sequences [10]. Biological databases, such as UniProt [48] and Gene Ontology (GO) [49], aggregate knowledge from experiments and studies, providing labeled data for GA. Knowledge graphs (KGs) further enrich GA by integrating heterogeneous biological data into structured formats [50]. Scientific publications serve as contextual evidence for gene annotations. Grounding annotations in traceable literature evidence ensures that final annotations are supported by verifiable sources.

Collaboration-related data capture how users engage with AI systems, for instance, the queries they enter, options they choose, and feedback they provide. These data help refine AI algorithms to better suit user needs.

Goals and tasks

Recent research in HAIC highlights that effective teamwork involves aligning AI behavior with human objectives, enabling both to contribute toward common goals [51]. These goals can be achieved through structured plans composed of subtasks. In hierarchical task analysis, a task is broken down into subtasks until a stop criterion is reached, often when the subtask consists of only a single operation [52]. For example, a single operation such as gene prediction may be performed by an AI tool, while others may be handled by humans, such as reviewing predictions. Plans define the sequence and structure of tasks, whether sequential or hierarchical, and help allocate specific subtasks to either human experts or AI systems [52].

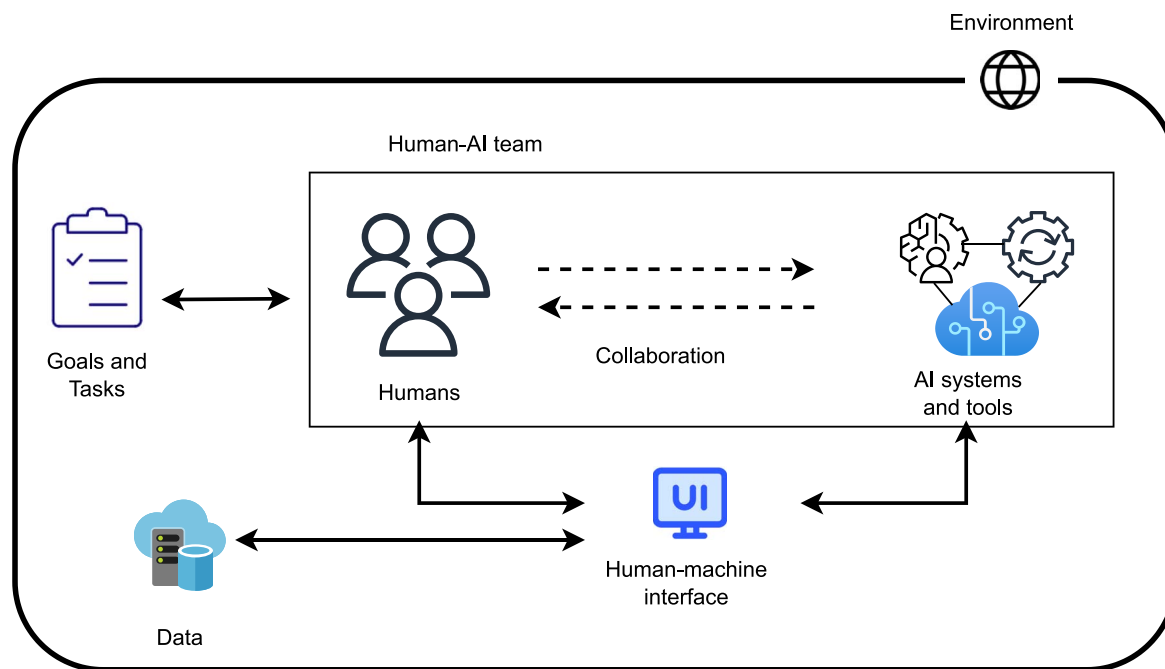


Figure 1. Key elements in human-AI collaboration for genome annotation.

Human-machine interface

The human-machine interface serves as the bridge enabling interaction between humans and AI systems. It includes various forms of interfaces, including command-line interfaces (CLIs), graphical user interfaces (GUIs), and conversational user interfaces (CUIs).

CLIs provide users with greater flexibility through programmatic access, customization options, and the ability to perform batch operations or access raw data directly. However, they can be challenging for users without programming experience. In contrast, GUIs provide a more user-friendly experience by allowing users to interact with AI systems through visual elements. GUIs are widely adopted in GA tools for tasks like visualizing feature locations, displaying evidence alignments, and presenting other relevant information. Genome browsers such as JBrowse [30] provide graphical interfaces for viewing GAs alongside supporting evidence tracks, while annotation platforms like Apollo [12] extend this functionality by enabling users to collaboratively edit and curate the annotations.

While GUIs are indispensable for the curation of structural annotations, CUIs backed by tool-wielding AI agents have the potential to fulfill an analogous role in functional annotation. Recent advancements in LLMs have significantly contributed to the growing interest in CUIs. CUIs allow users to interact with machines using natural language, making it easier to access information and perform tasks without needing to memorize complex commands or navigate intricate menus. For certain tasks, CUIs enhance HAIC by enabling effective and intuitive interactions between humans and AI [53].

Environment

Environment plays a critical role in shaping human perceptions of AI, influencing interaction dynamics, and affecting the AI system's capacity to interpret and respond to human input. It can be broadly categorized into digital, task, and team environments. The digital environment includes conditions and factors such as software platforms, interface design, and the availability of data for both humans and AI [54]. The task environment pertains

to the tasks that need to be completed, the constraints and limitations involved, and the desired outcomes [55]. The team environment refers to the dynamics and structures within a group of individuals (including both human and AI) working together [56]. It is characterized by the roles and relationships established among team members, communication patterns, and the level of cooperation and collaboration required to achieve overall goals.

Collaboration

Whether human-AI interaction constitutes true collaboration remains a subject of ongoing debate [57, 58]. In the HAICoGA framework, we use the term collaboration to refer to a structured interaction between human experts and AI systems, wherein complementary strengths are integrated to improve GA workflows. Effective HAIC relies on meaningful interaction and strategic alignment of distinct capabilities [20, 59]. Understanding these respective strengths is essential for designing effective collaborative systems in GA (Fig. 2; Supplementary Notes 3 and 4).

Human capabilities include abstract reasoning, situational awareness, and nuanced decision-making—capabilities that AI does not inherently possess, but may emulate to a limited extent through task conditioning and model fine-tuning. In contrast, AI offers scalable computation, rapid pattern recognition, and the efficient execution of repetitive or high-volume tasks such as candidate gene prioritization and evidence retrieval.

While automated GA tools can be error-prone, collaboration between humans and AI can improve the accuracy and reliability of the final outputs. AI systems should be designed to recognize tasks that exceed their confidence thresholds, such as annotating short genes or resolving complex duplications, and defer these cases to human curators for review. In turn, humans provide feedback by correcting errors, validating results, or suggesting alternative approaches. AI systems can then incorporate this feedback to refine their outputs, aligning more closely with human expectations and the dynamic context of the annotation process. This adaptive capability is referred to as contextual awareness. Bidirectional feedback mechanisms of this kind help prevent

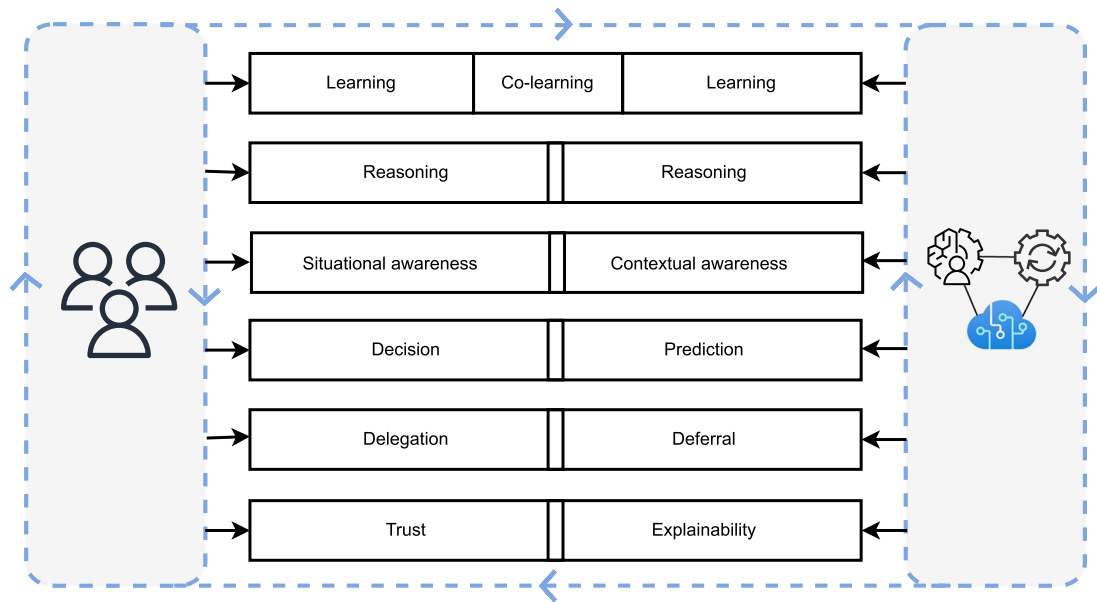


Figure 2. Key competencies for effective HAIC. Human competencies include learning, reasoning, situational awareness, decision-making, delegation, and trust. AI competencies mirror and complement these with learning, reasoning, contextual awareness, prediction, deferral, and explainability. Co-learning supports mutual adaptation. The dashed line represents an iterative feedback loop between humans and AI. See [Supplementary Note 4](#) for details.

erroneous predictions and reduces the risk of propagating errors in GAs. Moreover, ongoing efforts to improve AI explainability are critical to fostering effective collaboration, as they enhance transparency and build trust in automated systems.

AI agents bring opportunities to realize the HAICoGA framework

LLM-based AI assistants in biological and biomedical domains

There is an emerging trend of using LLM-backed AI assistants for research in biological and biomedical domains. These AI assistants can process human language inputs and generate responses that are coherent and contextually relevant within the interaction. We categorize these research based on the number of agents involved ([Table 1](#)). An agent refers to an AI system capable of interacting with humans or other agents and using tools to accomplish its tasks.

Some works in [Table 1](#), such as ChatNT, DRAGON-AI, and GeneGPT, are LLM-based models that take human language as input and generate direct answers without involving an agent. ChatNT is a multimodal AI system that integrates DNA, RNA, and protein sequences with neural language processing to solve various genomics tasks. ChatNT employs a modular LLM architecture that integrates a bidirectional DNA encoder with a unidirectional language decoder, enabling the interpretation of biological sequences in natural language. Similar architectures are discussed by Zhang *et al.* [60], who compare unidirectional and bidirectional LLMs in biological and chemical domains. Bidirectional models (e.g. BERT [61]) excel at encoding sequences for classification tasks, while unidirectional models (e.g. GPT [62]) are well-suited for generative tasks like text generation.

DRAGON-AI is a method that automatically generates ontology objects based on partial information from a user. All ontology terms and additional contextual information are translated into vector embeddings and indexed. Relevant contextual information is retrieved using a retrieval-augmented generation

(RAG) approach and added to construct a prompt, which is then passed as input to an LLM. The LLM completes the term object accordingly. GeneGPT uses few-shot learning to teach LLMs how to generate web APIs for accessing the National Center for Biotechnology Information (NCBI) databases and to answer biological questions based on the retrieved information. It handles both single-hop questions, which require a single API call, and more complex multi-hop questions that necessitate sequential API calls. For multi-hop questions, GeneGPT decomposes them into sub-questions, executing a chain of API calls to retrieve and integrate information step by step. While GeneGPT automates this process, its authors acknowledge that different types of errors are enriched in different tasks. Given the complexity inherent in multi-hop reasoning, such cases may still benefit from human oversight to ensure the accuracy and reliability of the results.

Phenomics Assistant builds an agent to call external tools based on user queries. It helps non-expert users query and interact with complex data from the Monarch Knowledge Graph. VarChat supports genetic professionals by providing concise summaries of scientific literature related to specific genomic variants. It interacts with external databases and utilizes user inputs to guide its querying and summarization processes. Both Phenomics Assistant and VarChat use a single-agent framework to provide a CUI that interacts with users and has the ability to use tools to solve user questions based on dynamic situations. The conversation history in the chat allows the agent to be aware of the user's state within tasks and incorporate feedback from external tools. Both systems also provide sources for the information in their responses, improving transparency in their processes.

Two-agent systems, ChatGSE, BioDiscoveryAgent, and Gene Agent, consist of a primary agent that interprets the user's query and selects appropriate tools to solve the problem, and a secondary agent that critically evaluates the results or verifies the factual accuracy of the output. The tools either retrieve and process information from various APIs to access online databases or scientific literature. The retrieved information is treated as a

Table 1. AI assistants in biological and biomedical domains

Method	Number of agents	Data	Task and goal	Team structure	Tool use	Explainability
ChatNT [63]	No agent	DNA, RNA, protein sequences and text data	Interpret biological information encoded in genome sequences and provide accurate predictions for various biological functions, such as gene expression prediction, DNA methylation, RNA stability and protein properties. Generate ontological terms.	NA	NA	NA
DRAGON-AI [64]	No agent	Structured data from existing ontologies and unstructured textual data from sources like GitHub issues		NA	NA	NA
GeneGPT [65]	No agent	Text data	Answer genomics-related questions by directly generating API request URLs to access and retrieve relevant biomedical information. Enhance accessibility to complex genomic information by enabling natural language querying of the Monarch knowledge graph.	NA	NA	NA
Phenomics Assistant [66]	Single AI agent	Monarch ^a knowledge graph	Support genetic professionals by providing concise summaries of scientific literature related to specific genomic variants. Answer user's questions using context from knowledge graphs and scientific papers; demonstrate the usability in cell type annotation task.	NA	Monarch Initiative API	The explainability of AI-generated answers by grounding them in data retrieved from the Monarch KG.
VarChat [67]	Single AI agent	Scientific literature and human genomic variants		NA	Query genomic databases; find and summarize the fragmented scientific literature	Informing users about the sources of its responses.
ChatGSE/ biochatter [68]	Two AI agents	Knowledge graph and scientific articles		Sequential	Information retrieval from knowledge graphs and the literature	Fact-checked and supplemented with context-specific information from documented sources.
BioDiscoveryAgent [69]	Two AI agents	Biological database (Reactome ^b 2022 database) and literature	Design genetic perturbation experiments that efficiently navigate the hypothesis space to identify a small subset of genes resulting in specific phenotypes. Generate biological process names for gene sets.	Sequential	Search the biomedical literature and execute code to analyze biological datasets	Detailed explanations for its choices, including citing relevant literature and detailing the reasoning behind selecting specific genes for perturbation.
GeneAgent [70]	Two AI agents	GO, Molecular Signature Database (MSigDB ^c), and a proteomics analysis system (NeST ^d)		Sequential	Call web APIs that connect to biological databases	Providing verification reports that detail the evidence supporting or refuting each generated name.
BRAD [71]	Multiple AI agents	Online literature repositories, Enrichr ^e and Gene Ontology databases	Automate bioinformatics workflows, enhancing the speed and efficacy of tasks such as gene enrichment analysis, literature searching, and running software pipelines.	Hierarchical	Search online literature, execute code to run software pipelines, such as enrichment analyze and visualization	Providing context-rich answers that include references to data sources and literature.

(continued)

Table 1. Continued

Method	Number of agents	Data	Task and goal	Team structure	Tool use	Explainability
BKGAgent [72]	Multiple AI agents	Clinical Knowledge and academic literature Graph	The primary task is Knowledge Graph Checking, which involves querying KGs, verifying the correctness of the information using external literature or databases, and identifying factual errors.	Hierarchical	Specific tools for interacting with knowledge graphs and scientific literature	Agent actions and decisions are traceable and justifiable, particularly in the context of verifying scientific claims and correcting knowledge graph data.
GenoAgent [73]	Multiple AI agents	GEO ^f and TCGA ^g databases	Automate the analysis of gene expression data to identify disease-associated genes.	Hierarchical	Various bioinformatics tools, such as those for data normalization and statistical analysis	Provide explainable results by documenting the decision-making process and the steps followed in the data analysis.
ProtAgents [74]	Multiple AI agents	Protein sequences, structural data, simulations and external databases	Automate and enhance the design of novel proteins with specific mechanical properties. This involves generating new proteins, analyzing their structures, and obtaining new first-principles data through physics simulations.	Hierarchical (dynamic, collaborative multi-agent environment)	Physics simulators and generative AI models, to perform tasks ranging from data retrieval to complex simulations of protein behaviors.	Provide explainable results by detailing the reasoning behind its decisions, the data used, and the methodologies applied.
[75]	Multiple AI agents	Single-cell RNA sequencing (scRNA-seq) data and literature	Replicate the experimental and analysis process of a scientific publication that explored gene expression relevant to SARS-CoV-2 entry into human cells. The goal is to validate the methods used in the original publication and to enhance the reproducibility and transparency of scientific research using AI.	Hierarchical	Software tools for data analysis and paper summary	Providing detailed breakdowns of its analytical processes and how decisions and analyses are derived, allowing for a transparent review of its methodology replication.
TAIS [76]	Multiple AI agents	TCGA, NCBI Gene and GEO databases	Identify disease-predictive genes from gene expression data.	Hierarchical	Various computational tools and methods integrated into the data processing and analysis workflows	Not detailed in the paper.
Virtual Lab [77]	Multiple AI agents	Public protein databases and SARS-CoV-2 variant data	Design and validate nanobody binders for SARS-CoV-2 variants using AI-driven workflows.	Hierarchical	Bioinformatics tools for protein analysis	Providing explainability by structuring AI-agent meetings, documenting decisions, and presenting clear computational workflows.

NA: not applicable. ^a<https://monarchinitiative.org/>, ^b<https://reactome.org/>, ^c<https://www.gsea-msigdb.org/gsea/msigdb>, ^d<https://idekerlab.ucsd.edu/nest/>, ^e<https://maayanlab.cloud/Enrichr/>, ^f<https://www.ncbi.nlm.nih.gov/geo/>, ^g<https://www.cancer.gov/ccg/research/genome-sequencing/tcga>

source to determine whether the answer is factually accurate compared with the original data. Keeping track of intermediate results from tools and the verification process enhances the agent's awareness of the current task status, potentially allowing it to adjust its actions accordingly in the next round of experiments. ChatGSE employs chain-of-thought reasoning to improve its problem-solving success. BioDiscoveryAgent follows the Reflection-Research Plan-Solution framework to enhance its reasoning capabilities. Both ChatGSE and BioDiscoveryAgent also incorporate self-verification mechanisms. These two agents operate in a sequential manner. All three systems provide some level of explainability by delivering context-rich answers that include references to data sources, literature, or verification reports. GeneAgent, which applies an AI agent for gene set enrichment analysis, focuses on autonomous interactions with domain-specific databases, followed by subsequent LLM verification.

Multi-agent systems are becoming increasingly popular for solving complex problems. These systems integrate multiple AI agents to automate and enhance critical workflows, significantly improving the speed and efficacy of tasks such as gene enrichment analysis, literature searches, and software pipeline executions. For instance, the BRAD system employs a hierarchical structure of agents to manage tasks like literature retrieval and enrichment analysis automation. These agents use a combination of in-context learning and a specialized planner to distribute and organize tasks efficiently. Another example is the BKGA-agent, which focuses on knowledge graph checking by querying knowledge graphs, verifying the accuracy of information through external literature or databases, and identifying factual discrepancies. The system's ability to dynamically query and cross-reference structured knowledge graphs and unstructured scientific texts illustrates the integration of RAG, ensuring relevance and contextual awareness throughout the information processing stages.

Similarly, GenoAgent and TAIS are tailored for analyzing gene expression data from sources like the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA). These systems leverage instruction learning and structured prompting to adapt their actions based on feedback and intermediate results, facilitating an iterative correction process that ensures the reliability and explainability of analytical outputs.

Beyond genomics, Virtual Lab exemplifies the application of multi-agent AI systems in experimental biomedical research. This system utilizes an AI-driven research framework, where a Principal Investigator AI leads a team of specialized agent, including a ML specialist, immunologist, and computational biologist, to design and validate nanobody binders for SARS-CoV-2 variants. The system's ability to document decision-making steps and optimize AI-driven workflows highlights the growing role of multi-agent systems in interdisciplinary research.

Lastly, ProtAgents showcases a multi-agent application in the design and analysis of novel proteins. By integrating real-time data from experiments and simulations, these agents can generate and analyze new proteins, adjusting their outputs based on dynamic inputs. The multi-agent system developed by Bersenev et al. [75] facilitates the replication of high-impact scientific studies by processing research papers and generating code to reproduce experiments, streamlining experimental validation and iterative scientific discovery.

Table 1 summarizes information from these studies, aligning certain elements with the HAIGoGA framework, including data,

tasks, goals, AI systems and tools, and team structure (environment). The data, tasks, goals, and tools are customized for different AI assistants. In studies involving multiple agents, these agents are often organized hierarchically, with a high-level agent (e.g. planner, leader, or manager) responsible for task distribution and coordination of the analysis process. Regarding the human-machine interface, three studies provide both GUI and CUI to facilitate human interaction with AI agents [66–68]. The most recent work, Virtual Lab [77], demonstrates the impact of HAIC through experiential evidence. In this framework, agents can defer tasks to other agents, as well as humans.

Cognitive functions, such as perception, reasoning, planning, and memory, are essential for enabling LLM-based agents to maintain contextual awareness and generate relevant responses in human-AI interactions. For example, the ReAct agent integrates reasoning and action, iteratively repeating this process until it determines a final response. The agent evaluates the current input along with past observations to decide the next step [78]. Some AI systems incorporate memory management to continuously track user interactions and dynamically recalibrate the agent's actions based on intermediate results and feedback [66–68, 71]. Table 1 shows that most studies support explainability through tracing agent actions, predictions, and the external data sources used.

Vision for the HAICoGA framework

Multi-agent system design in the HAICoGA framework

Through our review of current LLM agents in the biological and biomedical domains, we identified multi-agent systems as a promising approach for realizing the HAICoGA framework. Existing research primarily focuses on developing autonomous systems that minimize or even eliminate human intervention. However, such fully autonomous systems have demonstrated limited effectiveness in real-world applications [77, 79]. It is essential to keep humans in the loop to enhance system performance and reliability [18, 77].

Figure 3A illustrates an example of users collaborating with a multi-agent system to annotate gene functions. Based on the user's input query, the manager agent could use a method (e.g. ReAct) for breaking down the query into subtasks and assigns them to other agents according to their capabilities (Fig. 3C and D). The critique agent evaluates the quality of task results using metrics such as completeness, relevance, and other task-specific criteria, providing feedback and indicating the task's status. If necessary, agents can request additional input from the user. Once all tasks are completed, the manager agent compiles the final response and presents it to the user.

Building on the GA workflow described in the review by Ejigu et al. [10], we propose an automated GA agent along with several agents for manual annotation (categorized as manual curation agents in Fig. 3B, each assigned distinct roles, as detailed in Fig. 3D). While manual annotation is often performed based on the results of automated GA, newly added manual annotations can also enhance the automated GA system by providing additional gold-standard data, enabling continuous refinement of gene annotations.

Another key strength of multi-agent systems is that it allows for the internal refinement of answers. In the automated GA phase (Fig. 3C), the automated GA agent executes AI models and pipelines to perform specific tasks using genome data, such as predicting gene functions. The manager agent and critique agent

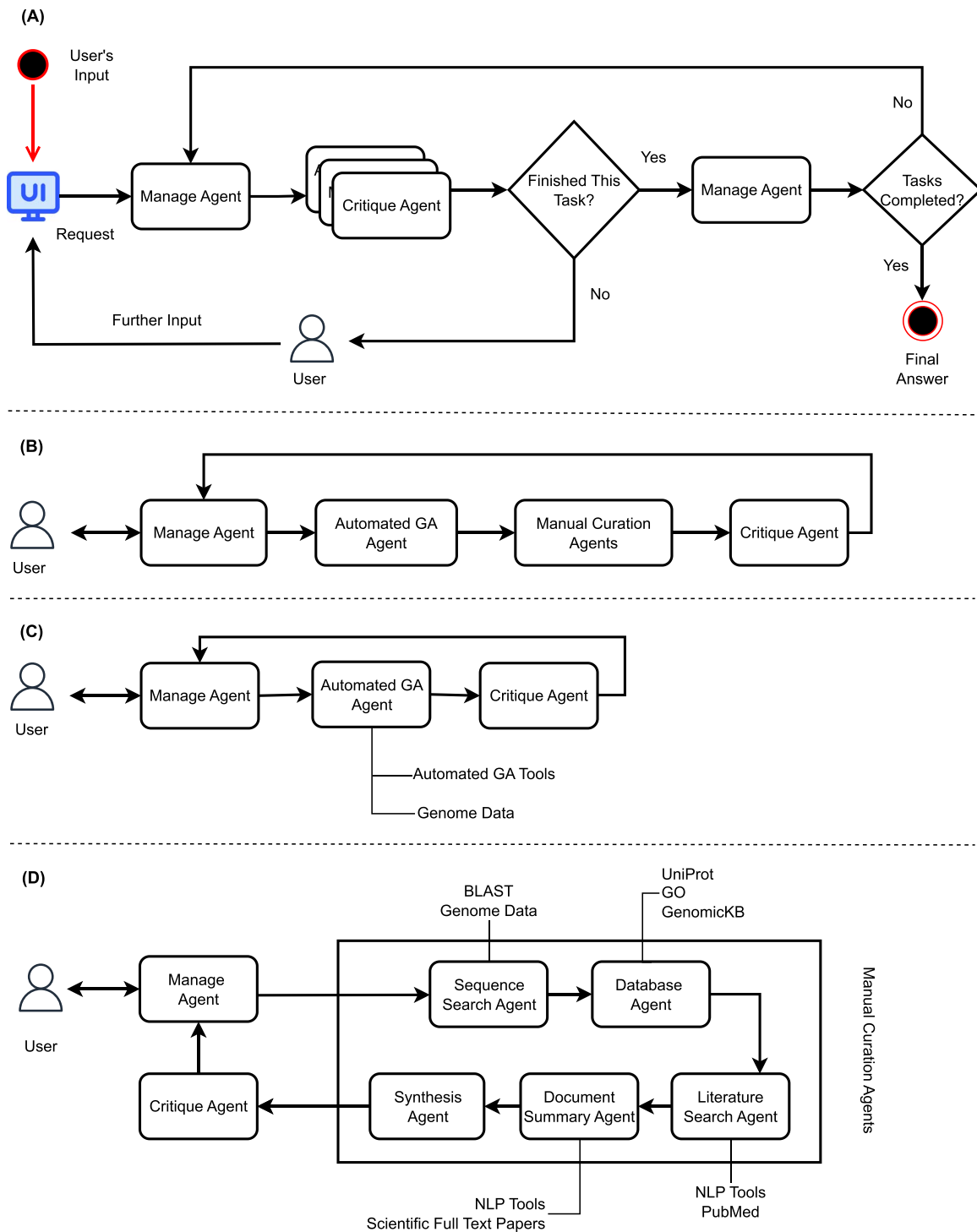


Figure 3. Multi-agent system design in the HAICoGA framework. (A) Overall multi-agent system design for human-AI collaborative GA. Users submit a GA query through an interactive user interface (UI). The UI requests the manager agent to analyze the task, decompose it into subtasks, and assign them to appropriate agents. While assisting with a subtask, an agent may request additional input from the user to complete the task successfully. The critique agent provides feedback on the outcomes, guiding the system's next steps. The manager agent monitors the global conversation history and intermediate results, updating the task plan as needed or finalizing the task and delivering the results to the user. (B) The top synergy layer of the multi-agent system designed for HAICoGA. Following the practical GA workflow [10], the multi-agent system consists of a user, a manager agent, an automated GA agent, multiple manual curation agents, and a critique agent. (C) Workflow of multi-agent collaboration in automated GA phase. The manager agent delegates the automated GA task to the automated GA agent, which manages a customized pipeline (or an AI model) using genome data to perform specific tasks. The critique agent analyzes the results, evaluates their quality, and suggests the next steps to the manager agent. This process can be repeated iteratively until the desired outcome is achieved. (D) Workflow of multi-agent collaboration in manual curation phase. A manual annotation process follows the automated GA phase. Due to the complexity of manual curation, the system includes several specialized agents performing distinct roles. The sequence search agent identifies homologous genes for a target gene, for example, by running BLAST against genome sequence data. The database agent retrieves gene function annotations from various databases. The literature search agent identifies relevant scientific papers for further analysis, while the document summarization agent extracts key information from these papers. The synthesis agent compiles all relevant data and submits it to the critique agent, which reviews the information and provides suggestions, such as whether the data is sufficient to address the user's query. Finally, the manager agent either updates the task plan or generates the final response.

contribute by summarizing results and providing feedback to the automated GA agent, which may prompt it to select alternative models or pipelines for gene function prediction. This iterative process enhances the quality of gene annotation. The self-improving loop continues until either the user or the manager agent decides to finalize the process and provide the final answer for the task.

The use of multiple agents also allows for specialization in the manual annotation system (Fig 3D), each assigned distinct attributes, including role, perception, and actions (tool use). These attributes enable agents to be optimized for specific domains or functions [80]. To manually annotate an uncharacterized gene, several guidelines recommend a workflow that involves using a tool (e.g. BLAST) to identify homologous proteins, retrieving functional annotations from existing databases and recent literature, and assigning these functions to the target protein [7, 33, 81]. Following these guidelines, the manager agent is responsible for designing this workflow and distributing tasks among specialized agents, including the sequence search agent, database agent, literature search agent, and document summary agent. The synthesis agent then aggregates the results, while the critique agent evaluates the output and provides feedback to the manager agent. Similar to the automated GA phase, the user could interrogate the results and refine prompts to continuously refine the quality of gene annotation.

Illustrative use cases of the HAICoGA framework

To demonstrate the practical use cases of the HAICoGA framework, we highlight the application of the GeneWhisperer system for gene annotation [82].

GeneWhisperer employs an LLM agent integrated with domain-specific tools to assist in generating functional hypotheses for genes, particularly uncharacterized genes in a reference genome. The system synthesizes multiple forms of evidence by identifying homologous proteins through sequence alignment, proposing relevant Gene Ontology (GO) terms based on functional similarity, and extracting gene-trait associations from scientific literature.

Following AI-assisted annotation, domain experts would review the generated hypotheses, validating them against species-specific literature and related annotations in other genomes. While experts do not generate annotations entirely from scratch, they are able to refine, correct, or reject AI-suggested annotations based on domain knowledge. As noted by Kudiabor et al. [83], AI-assisted annotations, particularly for novel genes, should not be considered definitive without supporting wet-lab experiments. Furthermore, we acknowledge that for certain genes, neither the user nor the AI system may be able to produce a meaningful annotation when no relevant information currently exists.

Another use case of the HAICoGA framework involves an AI assistant designed to improve consistency in gene function annotation. Manual curation often results in variability due to the difficulty in selecting standardized GO terms and corresponding Evidence and Conclusion Ontology (ECO) codes.

The AI assistant would analyze user-provided inputs, e.g. literature excerpts, and suggests appropriate GO and ECO terms. Users would review and refine these suggestions, maintaining expert oversight throughout the process. We demonstrated an example using ChatGPT as an LLM-based agent to assist in selecting GO and ECO terms (see [Supplementary Note 5](#)). While preliminary, this example illustrates the potential of general-purpose LLMs like ChatGPT can serve as accessible annotation

assistants. It also highlights the limitations of such models in domain-specific tasks, underscoring the need for future development of specialized AI assistants built on the HAICoGA framework.

These illustrative use cases demonstrate the practical viability of the HAICoGA framework in supporting GA tasks through synergistic human-AI workflows. Similar ideas have been implemented in other scientific domains. For example, the AI Co-Scientist system leverages a multi-agent architecture to collaborate with scientists in hypothesis generation, drug repurposing, and biomedical discovery [84]. This iterative collaboration between AI systems and domain experts reflects the same core principles underpinning HAICoGA.

By optimizing agents for specific annotation tasks and integrating expert feedback, HAICoGA aims to extend these advances into the genomics space. In the following section, we discuss the remaining challenges and technical considerations in building such systems.

Challenges in building the HAICoGA framework

Designing the architectural of a multi-agent system

The design of LLM-based multi-agent systems requires a modular and adaptive architecture in which specialized agents collaborate dynamically through structured interaction layers. These agents, each with distinct roles, leverage LLM capabilities for reasoning and task execution while interoperating with external resources such as datasets and tools to maintain contextual awareness. Achieving this requires balancing autonomy and alignment, as excessive autonomy may lead to goal deviations, whereas strict alignment can hinder adaptability [85]. Furthermore, managing dependencies among agents and ensuring scalability in resource usage are critical, especially as tasks grow more complex. Mechanisms for real-time adaptation and error correction are also essential to address inconsistencies and ensure robust, goal-oriented outcomes in complex environments. Finally, challenges remain in optimizing task allocation, fostering robust reasoning through iterative debates, managing complex contextual information, and enhancing memory management [86].

Developing novel ML/AI methods for enhancing HAIC

LLM agents, particularly unidirectional models, facilitate dynamic communication with users, but recent research highlights several critical challenges that may affect their collaborative effectiveness. Hallucination remains a significant concern, in which models generate plausible-sounding but factually unsupported content [87]. As such outputs can influence decision-making, they risk propagating false beliefs or even causing harm, underscoring the need for robust mitigation strategies. Fine-tuning bidirectional models with sufficient domain-specific training data can significantly improve their performance in tasks such as information extraction and classification. To further enhance reliability, systems could support continuous learning, enabling dynamic updates through human feedback and evolving contexts, as exemplified by reinforcement learning from human feedback [80].

Maintaining context over extended interactions is another area where LLMs often falter, leading to incoherent responses or an

inability to recall previous discussions. Vector databases offer a potential solution by enabling long-term memory management in LLM agents, allowing them to accumulate and organize memories over time. However, efficiently searching and retrieving relevant information from extensive memory stores remains challenging. Further advancements are needed to develop mechanisms for learning and updating metadata attributes across both procedural and semantic memory types [88]. MemGPT [89] exemplifies progress in this domain by intelligently managing different memory tiers to store and retrieve information effectively during long-term conversations.

Reasoning capabilities are pivotal for LLM agents to perform complex and nuanced tasks such as problem-solving, decision-making, and planning. Explicit reasoning steps not only improve task performance but also enhance model explainability and interpretability by providing rationales for predictions. While LLMs are primarily trained for next-token prediction, strategies like Chain of Thought (CoT) have demonstrated improvements in reasoning tasks by guiding models to articulate their reasoning explicitly. However, LLMs still face challenges in handling highly complex reasoning tasks or those involving subtle implicatures, necessitating ongoing research [90].

Requiring multidimensional evaluation methods to assess the HAICoGA workflow

Traditional GA evaluation metrics, such as coverage, precision, and accuracy, remain fundamental for assessing annotation quality [3, 91]. These measures indicate better outcomes when higher values are achieved; however, they provide relative rather than absolute benchmarks due to the absence of a comprehensive genome-wide gold standard. Many annotations remain provisional, relying on computational predictions or homologous transfers from model organisms.

In HAICoGA workflows, additional dimensions, such as explainability, are crucial for evaluation. Integrating orthologous information, along with detailed protein family and domain characterizations from diverse sources, enhances the explanatory depth and reliability of annotations [92]. Metrics that assess explanation generation and evidence quality are essential to ensuring the transparency of AI-assisted workflows. This aligns with frameworks for evaluating HAIC, which emphasize not only task success but also interaction quality, process dynamics, and ethical considerations [93].

Furthermore, optimizing the performance of human-AI teams requires a paradigm shift from individual AI optimization to assessing team-level outcomes. Evidence suggests that the most accurate AI system does not necessarily yield the best collaborative performance [94]. Effective collaboration depends on dynamic task allocation, mutual learning, and trust between human and AI agents. Metrics for evaluating such interactions must consider both qualitative factors, such as trust and satisfaction, and quantitative measures, such as decision impact and task completion time [93].

Adopting multi-dimensional evaluation frameworks, such as those emphasizing symbiotic HAIC modes, can provide holistic insights [93]. These frameworks should capture the dynamic, reciprocal nature of collaboration, extending beyond task success to evaluate how well humans and AI adapt to each other's strengths and limitations over time. Such comprehensive approaches are crucial for advancing the HAICoGA workflow and ensuring its alignment with both scientific rigor and practical utility.

Designing intuitive and interactive interfaces to facilitate HAIC

To investigate the challenges and opportunities in CUIs, we developed a chatbot prototype for curating information in gene functional annotation [82]. Additionally, we proposed applying conjoint analysis, a behavioral science method, to quantify the relative importance of four design features that influence users' trust in the system [95].

Initial testing of the prototype suggests that LLM agents have the potential to serve as valuable tools for collaborative GA when combined with human expertise. However, further research is needed to enhance their trustworthiness, particularly by improving explainability and providing confidence measures for AI-generated predictions [95].

To support these capabilities, future work will focus on integrating a dedicated graphical user interface (GUI) with the chatbot, particularly for structural annotation. Developing the right interface will be best served by taking a participatory or user-centered design approach and incorporating input from GA experts from the outset.

Risks and safeguards

The integration of LLM agents into scientific workflows introduces a set of risks that necessitate proactive and comprehensive mitigation strategies. The risks include the potential for generating misleading or harmful content, propagating biases, and compromising data privacy and security. These risks can arise from the inherent limitations of LLMs, such as their susceptibility to hallucination, over-reliance on training data, and the challenges of ensuring alignment with human values and ethical standards [96].

To mitigate these risks, a triadic framework involving human regulation, agent alignment, agent regulation and environmental feedback has been proposed [96]. Human regulation involves establishing clear guidelines and protocols for the responsible use and development of LLM agents in scientific contexts. This ensures ongoing human oversight and supports human-in-the-loop validation [97]. Agent alignment means that LLM agents are designed and trained to align with human intents and ethical standards, minimizing the risk of generating misleading or harmful content. Widely adopted safety mechanisms, such as those implemented in ChemCrow [18] and SafeScientist [98], can help ensure that agents operate within predefined boundaries and do not produce harmful outputs. Agent regulation and environmental feedback refer to the continuous monitoring and evaluation of LLM agent performance in real-world applications, enabling iterative refinement of their behavior. Feedback in multi-agent systems comes not only from human users but also from critique agents, external tools, and structured knowledge sources. Techniques like RAG exemplify how agents can be designed to incorporate trusted external knowledge sources, improving reliability and reducing the risk of hallucinated content [97].

Conclusion

In this paper, we first analyzed the pros and cons of automated GA methods and manual curation tools. We found that while automated GA methods generate annotations quickly, they have limitations, such as inaccurate gene predictions. On the other hand, manual curation can be highly accurate but requires intensive human labor and time. A human-AI collaborative GA approach

is necessary to leverage the strengths of both humans and AI, leading to more accurate and efficient GA.

Bringing together prior work in automated GA and manual curation, we then proposed the conceptual framework of HAICoGA. Our work bridges the gap between GA and human–AI collaborative communities, envisioning new possibilities in this multidisciplinary field. The emergence of LLM agents presents significant opportunities to realize HAICoGA workflows. However, many challenges and open questions remain in LLM agent research. The HAICoGA framework is still in its early stages of development, but it represents a step toward a comprehensive and efficient human–AI collaborative workflow for real-world applications in the future.

Glossary

Genome annotation (GA) is the process of identifying and characterizing functional elements within a genome, including genes, regulatory regions, and other biologically significant sequences. It involves the use of computational methods, such as machine learning (ML) and heuristic-based approaches, as well as manual curation by experts to improve accuracy. GA is essential for understanding gene functions, predicting protein structures, and exploring evolutionary relationships across species.

Artificial Intelligence (AI) refers to the simulation of human intelligence in machines, enabling them to perform tasks such as reasoning, learning, problem-solving, and decision-making. AI encompasses various techniques, including ML, deep learning, and natural language processing (NLP), to analyze complex data and automate decision-making. In GA, AI is used to enhance the efficiency of gene prediction, functional annotation, and data integration by processing large-scale biological datasets with minimal human intervention.

Machine Learning (ML) is a subset of AI that enables computers to learn patterns from data and make predictions or decisions without being explicitly programmed. In GA, ML algorithms are used to classify genes, predict functional elements, and enhance annotation accuracy by analyzing large-scale genomic datasets. ML approaches include supervised, unsupervised, and reinforcement learning, leveraging statistical models and neural networks to improve biological data interpretation.

Manual curation, also known as manual annotation, refers to the process in which human experts review, refine, and validate genome annotations to ensure accuracy and biological relevance. This process involves analyzing computationally generated annotations, resolving ambiguities, and incorporating insights from experimental data and scientific literature.

Human–AI collaboration (HAIC) refers to the dynamic interaction between humans and AI systems, where both work together toward overall goals by leveraging their complementary strengths. Unlike automation, where AI operates independently, or augmentation, where AI enhances human capabilities, HAIC involves a continuous exchange of information, decision-making, and adaptation over time.

Knowledge graphs (KGs) are structured representations of relationships between biological entities, such as variants, genes, proteins, pathways, phenotypes, and diseases. They encode known interactions and associations in a graph format, where nodes represent entities and edges denote relationships. KGs facilitate data integration, reasoning, and discovery in genomics by linking heterogeneous biological information sources.

Large language models (LLMs) are AI models trained on massive datasets of text and code. They can generate human-quality text,

translate languages, follow user instructions for task procedures [99, 100], use external tools [101], and answer user questions based on specific contexts [102]. A common architectural foundation for LLMs is the *Transformer* [103], which enables efficient modeling of long-range dependencies in sequences through self-attention mechanisms. Variations of this architecture include encoder-only models (e.g. BERT [61]), decoder-only models (e.g. GPT [62]), and encoder–decoder hybrids (e.g. T5 [104]). These architectures may be *bidirectional*, capturing context from both preceding and following tokens (as in BERT), or *unidirectional*, processing text left-to-right to generate coherent outputs (as in GPT models).

AI agent is an autonomous or semi-autonomous entity within a multi-agent system that performs specific tasks, interacts with other agents, and operates based on predefined rules, learned behaviors, or external inputs. Agents may specialize in different roles, such as task management, data retrieval, reasoning, or quality assessment, and they communicate within structured frameworks to enhance human–AI collaboration.

Key Points

- While genome annotation is complex and challenging, heavy reliance on automated methods can introduce errors.
- Manual curation is necessary for accurate annotations but requires significant time and effort.
- Our novel contribution is HAICoGA, the first conceptual framework for human–AI collaborative genome annotation.
- We further present an example of HAICoGA framework and future research directions to realize this framework.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Competing interests

The authors declare no competing interests.

Funding

None declared.

Data availability

All data presented in tables and figures were compiled from publicly available sources, which are cited in the manuscript and supplementary materials.

References

1. Eraslan G, Avsec Ž, Gagneur J. et al. Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* 2019;**20**:389–403. <https://doi.org/10.1038/s41576-019-0122-6>
2. Zou J, Huss M, Abid A. et al. A primer on deep learning in genomics. *Nat Genet* 2019;**51**:12–8. <https://doi.org/10.1038/s41588-018-0295-5>

3. Mahood EH, Kruse LH, Moghe GD. Machine learning: a powerful tool for gene function prediction in plants. *Appl Plant Sci* 2020;**8**:e11376. <https://doi.org/10.1002/aps3.11376>
4. Altschul SF, Gish W, Miller W. et al. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
5. Stanke M, Keller O, Gunduz I. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* 2006;**34**:W435–9.
6. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* 2012;**13**:329–42. <https://doi.org/10.1038/nrg3174>
7. Madupu R, Brinkac LM, Harrow J. et al. Meeting report: a workshop on best practices in genome annotation. *Database* 2010, 2010:baq001.
8. Tatusova T, DiCuccio M, Badretdin A. et al. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res* 2016;**44**:6614–24. <https://doi.org/10.1093/nar/gkw569>
9. Zerbino DR, Frankish A, Flicek P. Progress, challenges, and surprises in annotating the human genome. *Annu Rev Genomics Hum Genet* 2020;**21**:55–79. <https://doi.org/10.1146/annurev-genom-121119-083418>
10. Ejigu GF, Jung J. Review on the computational genome annotation of sequences obtained by next-generation sequencing. *Biology* 2020;**9**:295.
11. Tello-Ruiz MK, Marco CF, Hsu F-M. et al. Double triage to identify poorly annotated genes in maize: the missing link in community curation. *PLoS One* 2019;**14**:e0224086. <https://doi.org/10.1371/journal.pone.0224086>
12. Lee E, Helt GA, Reese JT. et al. Web Apollo: a web-based genomic annotation editing platform. *Genome Biol* 2013;**14**:1–13:R93. <https://doi.org/10.1186/gb-2013-14-8-r93>
13. Haas BJ, Wortman JR, Ronning CM. et al. Complete reannotation of the arabidopsis genome: methods, tools, protocols and the final release. *BMC Biol* 2005;**3**:1–19.
14. Aodha O, M, Stathopoulos V, Brostow GJ. et al. Putting the scientist in the loop—accelerating scientific progress with interactive machine learning. In: *2014 22nd International Conference on Pattern Recognition*, Stockholm, Sweden, 2014, pp. 9–17, IEEE, Piscataway, NJ, USA.
15. Gao S, Fang A, Huang Y. et al. Empowering biomedical discovery with AI agents. *arXiv preprint arXiv:2404.02831*. 2024.
16. van der Wal D, Jhun I, Laklout I. et al. Biological data annotation via a human-augmenting AI-based labeling system. *NPJ Digital Medicine* 2021;**4**:145. <https://doi.org/10.1038/s41746-021-00520-6>
17. Tschandl P, Rinner C, Apalla Z. et al. Human–computer collaboration for skin cancer recognition. *Nat Med* 2020;**26**:1229–34. <https://doi.org/10.1038/s41591-020-0942-0>
18. Bran AM, Cox S, Schilter O. et al. Augmenting large language models with chemistry tools. *Nat Mach Intell* 2024;**6**:525–35. <https://doi.org/10.1038/s42256-024-00832-8>
19. James Wilson H, Daugherty PR. Collaborative intelligence: humans and ai are joining forces. *Harv Bus Rev* 2018;**96**:114–23.
20. Schleiger E, Mason C, Naughtin C. et al. Collaborative intelligence: a scoping review of current applications. *Appl Artif Intell* 2024;**38**:2327890. <https://doi.org/10.1080/08839514.2024.2327890>
21. Goldberg K. Robots and the return to collaborative intelligence. *Nat Mach Intell* 2019;**1**:2–4. <https://doi.org/10.1038/s42256-018-0008-x>
22. Huang M-H, Rust RT. A framework for collaborative artificial intelligence in marketing. *J Retailing* 2022;**98**:209–23. <https://doi.org/10.1016/j.jretai.2021.03.001>
23. Reed JL, Famili I, Thiele I. et al. Towards multidimensional genome annotation. *Nat Rev Genet* 2006;**7**:130–41. <https://doi.org/10.1038/nrg1769>
24. Kyrpides NC. Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nat Biotechnol* 2009;**27**:627–32. <https://doi.org/10.1038/nbt.1552>
25. Hall M, van der Maaten L, Gustafson L. et al. A systematic study of bias amplification. *arXiv preprint arXiv:2201.11706*. 2022.
26. Cheng C-Y, Krishnakumar V, Chan AP. et al. Araport11: a complete reannotation of the arabidopsis thaliana reference genome. *Plant J* 2017;**89**:789–804. <https://doi.org/10.1111/tpj.13415>
27. Lamesch P, Berardini TZ, Li D. et al. The arabidopsis information resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 2012;**40**:D1202–10. <https://doi.org/10.1093/nar/gkr1090>
28. Lewis SE, Searle SMJ, Harris N. et al. Apollo: a sequence annotation editor. *Genome Biol* 2002;**3**:1–14.
29. Dunn NA, Unni DR, Diesh C. et al. Apollo: democratizing genome annotation. *PLoS Comput Biol* 2019;**15**:e1006790. <https://doi.org/10.1371/journal.pcbi.1006790>
30. Buels R, Yao E, Diesh CM. et al. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol* 2016;**17**:1–12.
31. Wei C-H, Allot A, Leaman R. et al. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res* 2019;**47**:W587–93. <https://doi.org/10.1093/nar/gkz389>
32. Mao Y, Van Auken K, Li D. et al. Overview of the gene ontology task at BioCreative IV. *Database* 2014, 2014:bau086.
33. Drabkin HJ, Blake JA, Mouse Genome Informatics Database. Manual gene ontology annotation workflow at the mouse genome informatics database. *Database* 2012, 2012:bas045.
34. Chhetri MB, Tariq S, Singh R. et al. Towards human-AI teaming to mitigate alert fatigue in security operations centres. *ACM Trans Internet Technol* 2024;**24**:1–22. <https://doi.org/10.1145/3670009>
35. Hartikainen M, Spurava G, Väänänen K. Human-AI collaboration in smart manufacturing: key concepts and framework for design. *HHA1 2024: Hybrid Human AI Systems for the Social Good* 162–172:2024.
36. Martin A V-I, Selva D. A framework to study human-ai collaborative design space exploration. In: *Proceedings of the ASME 2021 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Virtual, Online, 2021, p. V006T06A052. American Society of Mechanical Engineers, New York, NY, USA.
37. Dubey A, Abhinav K, Jain S. et al. HACO: a framework for developing human-AI teaming. In: *Proceedings of the 13th Innovations in Software Engineering Conference on Formerly known as India Software Engineering Conference*, Jabalpur, India, 2020, pp. 1–9, Association for Computing Machinery, New York, NY, USA.
38. Dellermann D, Calma A, Lipusch N. et al. The future of human-AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. *arXiv preprint arXiv:2105.03354*. 2021.
39. Cooper N. Cognitive biases. In: Cooper N, Frain J, eds. *ABC of clinical reasoning*. Chichester, UK: Wiley-Blackwell, 2017, pp. 41–46.
40. Klein G. Naturalistic decision making. *Hum Factors* 2008;**50**:456–60. <https://doi.org/10.1518/001872008X288385>
41. Kahneman, Slovic P, and Tversky A. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge, UK, 1982, <https://doi.org/10.1017/CBO9780511809477>.

42. Simon HA, Simon HA. Scientific discovery and the psychology of problem solving. In *Models of Discovery: And Other Topics in the Methods of Science*, pp. 286–303. Springer, Dordrecht, Netherlands, 1977, https://doi.org/10.1007/978-94-010-9521-1_16.
43. Abels A, Lenaerts T. Wisdom from diversity: bias mitigation through hybrid human-llm crowds. arXiv preprint arXiv:2505.12349. 2025.
44. Van Auken K, Schaeffer ML, McQuilton P. et al. BC4GO: a full-text corpus for the BioCreative IV GO task. *Database* 2014, 2014:bau074.
45. Koonin EV, Galperin MY, Koonin EV. et al. Genome annotation and analysis. *Sequence—evolution—function: computational approaches in comparative Genomics*, Springer; Boston, MA, 2003, pp. 193–226. https://doi.org/10.1007/978-1-4757-3783-7_6
46. Paris C, Reeson A. *What's the Secret to Making Sure AI Does Not Steal your Job? Work with it, Not against it*. Baltimore: Johns Hopkins University Press, 2024. 177–81.
47. Mungall CJ, Misra S, Berman BP. et al. An integrated computational pipeline and database to support whole-genome sequence annotation. *Genome Biol* 2002;**3**:1–11. <https://doi.org/10.1186/gb-2002-3-12-research0081>
48. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Res* 2023;**51**:D523–31. <https://doi.org/10.1093/nar/gkac1052>
49. Gene Ontology Consortium. The gene ontology resource: enriching a gold mine. *Nucleic Acids Res* 2021;**49**:D325–34. <https://doi.org/10.1093/nar/gkaa1113>
50. Sunil RS, Lim SC, Itharajula M. et al. The gene function prediction challenge: large language models and knowledge graphs to the rescue. *Curr Opin Plant Biol* 2024;**82**:102665. <https://doi.org/10.1016/j.pbi.2024.102665>
51. Lou B, Lu T, Raghu TS. et al. Unraveling human-AI teaming: a review and outlook. arXiv preprint arXiv:2504.05755. 2025.
52. Bligård L-O, Osvalder A-L. CCPE: methodology for a combined evaluation of cognitive and physical ergonomics in the interaction between human and machine. *Hum Factors Ergon Manuf Serv Ind* 2014;**24**:685–711. <https://doi.org/10.1002/hfm.20512>
53. Feng X, Chen Z-Y, Qin Y. et al. Large language model-based human-agent collaboration for complex task solving. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*, Miami, Florida, USA, Association for Computational Linguistics, Stroudsburg, PA, USA, 2024, pp. 1336–1357.
54. Madni AM, Madni CC. Architectural framework for exploring adaptive human-machine teaming options in simulated dynamic environments. *Systems* 2018;**6**:44. <https://doi.org/10.3390/systems6040044>
55. National Academies of Sciences, Engineering, and Medicine. *Human-AI Teaming: State-of-the-Art and Research Needs*. Washington, DC: The National Academies Press, 2022, pp. 11–18. <https://doi.org/10.17226/26355>.
56. Salas E, Shuffler ML, Thayer AL. et al. Understanding and improving teamwork in organizations: a scientifically based practical guide. *Hum Resour Manage* 2015;**54**:599–622. <https://doi.org/10.1002/hrm.21628>
57. Schmutz JB, Outland N, Kerstan S. et al. Ai-teaming: redefining collaboration in the digital era. *Curr Opin Psychol* 2024; 58:101837.
58. O'Neill TA, McNeese NJ, Barron A. et al. Human-autonomy teaming: a review and analysis of the empirical literature. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 2022;**64**:904–38. <https://doi.org/10.1177/0018720820960865>
59. Lyons JB, Sycara K, Lewis M. et al. Human-autonomy teaming: Definitions, debates, and directions. *Front Psychol* 2021;**12**:589585. <https://doi.org/10.3389/fpsyg.2021.589585>
60. Zhang Q, Ding K, Lyv T. et al. Scientific large language models: a survey on biological & chemical domains. *ACM Computing Surveys*. 2025;**57**:1–38.
61. Devlin J, Chang M-W, Lee K. et al. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, eds. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies*, Minneapolis, MN, USA, Association for Computational Linguistics, Stroudsburg, PA, USA, volume 1 (long and short papers), pp. 4171–86, 2019.
62. Radford A, Jeffrey W, Child R. et al. Language models are unsupervised multitask learners. *OpenAI blog* 2019;**1**:9.
63. Richard G, de Almeida BP, Dalla-Torre H. et al. ChatNT: a multimodal conversational agent for DNA, RNA and protein tasks, bioRxiv preprint bioRxiv:2024.04.30.591835. 2024
64. Toro S, Anagnostopoulos AV, Bello S. et al. Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI). *Journal of Biomedical Semantics*. 2024;**15**:19.
65. Jin Q, Yang Y, Chen Q. et al. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics* 2024;**40**:btac075.
66. Neil STO', Schaper K, Elsarboukh G. et al. Phenomics Assistant: an interface for LLM-based biomedical knowledge graph exploration bioRxiv preprint bioRxiv:2024.01.31.578275. 2024
67. De Paoli F, Berardelli S, Limongelli I. et al. VarChat: the generative ai assistant for the interpretation of human genomic variations. *Bioinformatics* 2024;**40**:btac183.
68. Lobentanzer S, Feng S., The BioChatter Consortium. et al. A platform for the biomedical application of large language models. *Nature biotechnology* 2025;**43**:166–9.
69. Roohani Y, Vora J, Huang Q. et al. BioDiscoveryAgent: an AI agent for designing genetic perturbation experiments. arXiv preprint arXiv:2405.17631. 2024.
70. Wang Z, Jin Q, Wei C-H. et al. GeneAgent: self-verification language agent for gene set knowledge discovery using domain databases. arXiv preprint arXiv:2405.16205. 2024.
71. Pickard J, Prakash R, Choi MA. et al. Automatic biomarker discovery and enrichment with BRAD. *Bioinformatics* 2025;**41**:btac159.
72. Lin X, Ma S, Shan J. et al. BioKGBench: a knowledge graph checking benchmark of AI agent for biomedical science. arXiv preprint arXiv:2407.00466. 2024.
73. Liu H, Wang H. GenoTEX: a benchmark for evaluating LLM-based exploration of gene expression data in alignment with bioinformaticians. arXiv preprint arXiv:2406.15341. 2024.
74. Ghafarollahi A, Buehler MJ. ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digital Discovery* 2024;**3**: 1389–409.
75. Bersenev D, Yachie A, Palaniappan SK. Replicating a high-impact scientific publication using systems of large language models. bioRxiv preprint bioRxiv:2024.04.08.588614. 2024
76. Liu H, Li Y, Jian J. et al. Toward a team of AI-made scientists for scientific discovery from gene expression data. arXiv preprint arXiv:2402.12391. 2024.
77. Swanson K, Wu W, Bulaong NL. et al. The virtual lab: AI agents design new SARS-COV-2 nanobodies with experimental validation. bioRxiv preprint bioRxiv:2024.11.11.623004. 2024
78. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, and Cao Y. ReAct: synergizing reasoning and acting in language

- models. In *International Conference on Learning Representations (ICLR) Workshop on Foundation Models for Decision Making*, Kigali, Rwanda, 2023.
79. Liu X, Yu H, Zhang H. et al. AgentBench: evaluating LLMs as agents. arXiv preprint arXiv:2308.03688. 2023.
 80. Xi Z, Chen W, Guo X. et al. The rise and potential of large language model based agents: a survey. *Science China Information Sciences*. 2025;**68**:121101.
 81. McDonnell E, Strasser K, and Tsang A. Manual gene curation and functional annotation. In: de Vries RP, Tsang A, Grigoriev IV, eds., *Fungal Genomics: Methods and Protocols*, pp. 185–208. Springer, New York, NY, 2018, https://doi.org/10.1007/978-1-4939-7804-5_16.
 82. Li X, Whan A, McNeil M. et al. GeneWhisperer: enhancing manual genome annotation with large language models bioRxiv preprint bioRxiv:2025.03.30.646211. 2025
 83. Kudiabor H. Virtual lab powered by 'AI scientists' supercharges biomedical research. *Nature* 2024;**636**:532–3. <https://doi.org/10.1038/d41586-024-01684-3>
 84. Gottweis J, Weng W-H, Daryin A. et al. Towards an ai co-scientist. arXiv preprint arXiv:2502.18864. 2025.
 85. Händler T. Balancing autonomy and alignment: a multi-dimensional taxonomy for autonomous LLM-powered multi-agent architectures. arXiv preprint arXiv:2310.03659. 2023.
 86. He J, Treude C, Lo D. LLM-based multi-agent systems for software engineering: vision and the road ahead. *ACM Transactions on Software Engineering and Methodology*. 2025;**34**:1–30.
 87. Huang L, Yu W, Ma W. et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*. 2025;**43**: 1–55.
 88. Hatalis K, Christou D, Myers J. et al. Memory matters: the need to improve long-term memory in LLM-agents. In *Proceedings of the AAAI Symposium Series*, AAAI Press, Washington, DC, USA, 2023;**2**:277–80. <https://doi.org/10.1609/aaais.v2i1.27688>
 89. Packer C, Wooders S, Lin K. et al. MemGPT: towards LLMs as operating systems. arXiv preprint arXiv:2310.08560. 2023.
 90. Huang J, Chang KC-C. Towards reasoning in large language models: a survey. In *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada, Association for Computational Linguistics, Stroudsburg, PA, pp. 1049–1065.
 91. Ouzounis CA, Karp PD. The past, present and future of genome-wide re-annotation. *Genome Biol* 2002;**3**:comment2001.1–6. <https://doi.org/10.1186/gb-2002-3-2-comment2001>
 92. Kirilenko BM, Munegowda C, Osipova E. et al. Integrating gene annotation with orthology inference at scale. *Science* 2023;**380**:eabn3107. <https://doi.org/10.1126/science.abn3107>
 93. Fragiadakis G, Diou C, Kousiouris G. et al. Evaluating human-AI collaboration: a review and methodological framework. arXiv preprint arXiv:2407.19098. 2024.
 94. Bansal G, Nushi B, Kamar E. et al. Is the most accurate ai the best teammate? Optimizing AI for teamwork. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI Press, Palo Alto, California USA, Vol. **35**, pp. 11405–14, 2021.
 95. McGrath M, Cooper P, Duenser A. et al. A novel method for trust-sensitive design: applying conjoint analysis to machine-assisted genome annotation. In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery, New York, NY, USA, Vol. **1–9**, p. 2025.
 96. Tang X, Jin Q, Zhu K. et al. Prioritizing safeguarding over autonomy: risks of LLM agents for science. arXiv preprint arXiv:2402.04247. 2024.
 97. Huang K, Yuanhao Q, Cousins H. et al. CRISPR-GPT: an LLM agent for automated design of gene-editing experiments. arXiv preprint arXiv:240418021 2024.
 98. Zhu K, Zhang J, Qi Z. et al. SafeScientist: toward risk-aware scientific discoveries by LLM agents. arXiv preprint arXiv:2505.23559. 2025.
 99. Wen H, Li Y, Liu G. et al. Empowering LLM to use smartphone for intelligent task automation. arXiv preprint arXiv:2308.15272. 2023.
 100. Singh I, Blukis V, Mousavian A. et al. ProgPrompt: generating situated robot task plans using large language models. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11523–30. IEEE, London, UK, 2023.
 101. Schick T, Dwivedi-Yu J, Dessì R. et al. Toolformer: language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761. 2023.
 102. Brown T, Mann B, Ryder N. et al. Language models are few-shot learners. *Advances in neural information processing systems*, Curran Associates Inc., Red Hook, NY, USA, 2020;**33**: 1877–901.
 103. Vaswani A, Shazeer N, Parmar N. et al. Attention is all you need. *Advances in neural information processing systems*, Curran Associates Inc., Red Hook, NY, USA, 2017;**30**.
 104. Raffel C, Shazeer N, Roberts A. et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J Mach Learn Res* 2020;**21**:1–67.