OXFORD UNIVERSITY PRESS | Briefings In Bioinformatics

A Conceptual Framework for Human-AI Collaborative Genome Annotation

Journal:	*Briefings in Bioinformatics*
Manuscript ID	BIB-25-0517.R1
Manuscript Type:	Review
Date Submitted by the Author:	n/a
Complete List of Authors:	Li, Xiaomei; CSIRO, Agriculture and Food Whan, Alex; CSIRO, Agriculture & Food Mcneil, Meredith; CSIRO, Agriculture & Food Starns, David; University of Leeds, School of Molecular and Cellular Biology Irons, Jessica; CSIRO Data61 Business Unit Andrew, Samuel; CSIRO Suchecki, Rad; Alkahest Inc
Keywords:	Genome annotation, human, artificial intelligence, collaboration, conceptual framework, large language model
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
main.tex LaTeX_Suppl_Files.7z	

Briefings in Bioinformatics, 2023, pp. 1–25

doi: DOI HERE

Advance Access Publication Date: Day Month Year

Review

REVIEW

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] and Rad Suchecki[6,7,*]

[1]AGRICULTURE & FOOD, CSIRO, 26 Pembroke Rd, 2122, NSW, Australia, [2]AGRICULTURE & FOOD, CSIRO, 2-40 Clunies Ross Street, Acton, 2601, ACT, Australia, [3]AGRICULTURE & FOOD, CSIRO, 306 Carmody Road, St Lucia, 4067, QLD, Australia, [4]School of Molecular and Cellular Biology, University of Leeds, Woodhouse Lane, LS2 9JT, Leeds, United Kingdom, [5]Data 61, CSIRO, 13 Garden Street, Eveleigh, 2015, NSW, Australia, [6]AGRICULTURE & FOOD, CSIRO, Waite Road, Urrbrae, 5064, SA, Australia and [7], Present address: Alkahest Inc., 125 Shoreway Rd D, San Carlos, 94070, CA, United States

*Corresponding author. Maisie.Li@csiro.au, rsuchecki@alkahest.com.

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

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 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), which leverages the synergistic partnership between humans and artificial intelligence to enhance human capabilities and accelerate the genome annotation process. Additionally, we explore the potential of integrating Large Language Models (LLMs) 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.

Key words: 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 is 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 genome annotation [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]. 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 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 machine learning 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 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 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 genome annotation 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, 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 multi-dimensional, spanning from the nucleotide level to the biological system level [23]. Genomic elements of interest include, but are not limited to, single nucleotide polymorphisms (SNPs), coding genes, non-coding genes, regulatory elements and other non-coding 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 faces serious limitations and challenges (Supplementary Note 1).

Due to the limitations of current computational tools, automated genome annotation (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 machine learning 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 genome annotations. 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 labour-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 [32] assist in extracting biological entities and gene functions from literature, yet manual curation remains challenging. This is evident in efforts like BioCreative IV [33], which highlight the expertise and validation required to curate functions even with automated support [34].

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 [35]. 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 human-AI collaboration in GA (Level 2).

The emergence of human-AI collaboration (HAIC) frameworks and taxonomies across various domains has advanced our understanding of effective human-AI collaboration. For instance, Hartikainen *et al.* [36] propose a framework for designing collaborative systems in smart manufacturing, while Viros i Martin *et al.* [37] present a model for human-AI co-design in Design Space Exploration. Dubey *et al.* [38] introduce a framework for successful human-AI teaming in contact centers. Despite this progress, applying HAIC to genome annotation presents unique challenges. We build on these frameworks to introduce HAICoGA and propose future research directions for collaborative genome annotation.

A conceptual framework of human-AI collaborative genome annotation

To support collaborative genome annotation, HAICoGA incorporates and extends key concepts from general HAIC theories [36, 38, 39]. As shown in Figure 1, it includes seven key elements: humans, AI systems and tools, data, goals and tasks, human-machine interface, environment, and collaboration.

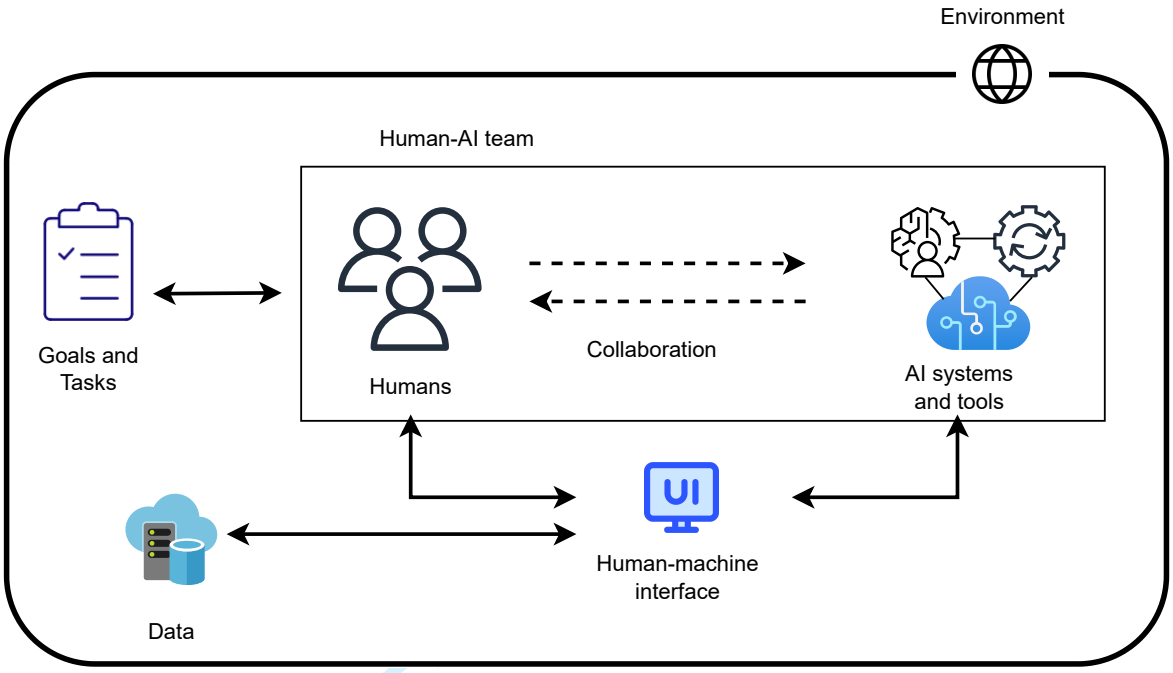

Fig. 1. Key elements in human-AI collaborative genome annotation. Humans and AI work together as a team to perceive the environment in which they operate. To achieve the high-level goal, they decompose the task into sub-tasks and objectives. Through the human-machine interface, humans and AI utilize the available data to carry out tasks and transition to a new state within the environment. This state may lead to updates in the list of tasks and goals or the addition of new data until the final goal is achieved. The collaboration between humans and AI is dynamic, allowing them to perform individual tasks independently while collaborating on shared tasks when necessary.

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 [40]. Research suggests experts rely heavily on their previous experience on the task, as well as their deep and often tacit knowledge of the domain [41]. Decision-makers often employ mental shortcuts and heuristics that are efficient but can be prone to cognitive bias [42, 43].

Collaborating with LLMs has the potential to mitigate individual biases when the complementary strengths of human judgment and machine predictions are effectively leveraged [44]. Still, bias in human curation remains an open and ongoing challenge in the biological domain. For example, the BC4GO study [45] 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 [41]. 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 [46]. Consequently, while AI can assist and augment decision-making, the role of experienced human curators remains indispensable [47].

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, 48]. 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 human-AI collaboration. 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 [49] and Gene Ontology (GO) [50], 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 [51]. 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 human-AI collaboration (HAIC) highlights that effective teamwork involves aligning AI behavior with human objectives, enabling both to contribute toward common goals [52]. 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 [53]. 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 [53].

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 genome annotation 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 genome

annotations 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 human-AI collaboration by enabling effective and intuitive interactions between humans and AI [54].

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 [55]. The task environment pertains to the tasks that need to be completed, the constraints and limitations involved, and the desired outcomes [56]. The team environment refers to the dynamics and structures within a group of individuals (including both human and AI) working together [57]. 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 [58, 59]. 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 genome annotation workflows. Effective human-AI collaboration relies on meaningful interaction and strategic alignment of distinct capabilities [20, 60]. Understanding these respective strengths is essential for designing effective collaborative systems in genome annotation (Figure 2; Supplementary Notes 3-4).

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. Bi-directional feedback mechanisms of this kind help prevent erroneous predictions and reduces the risk of propagating errors in genome annotations. Moreover, ongoing efforts to improve AI explainability are critical to fostering effective collaboration, as they enhance transparency and build trust in automated systems.

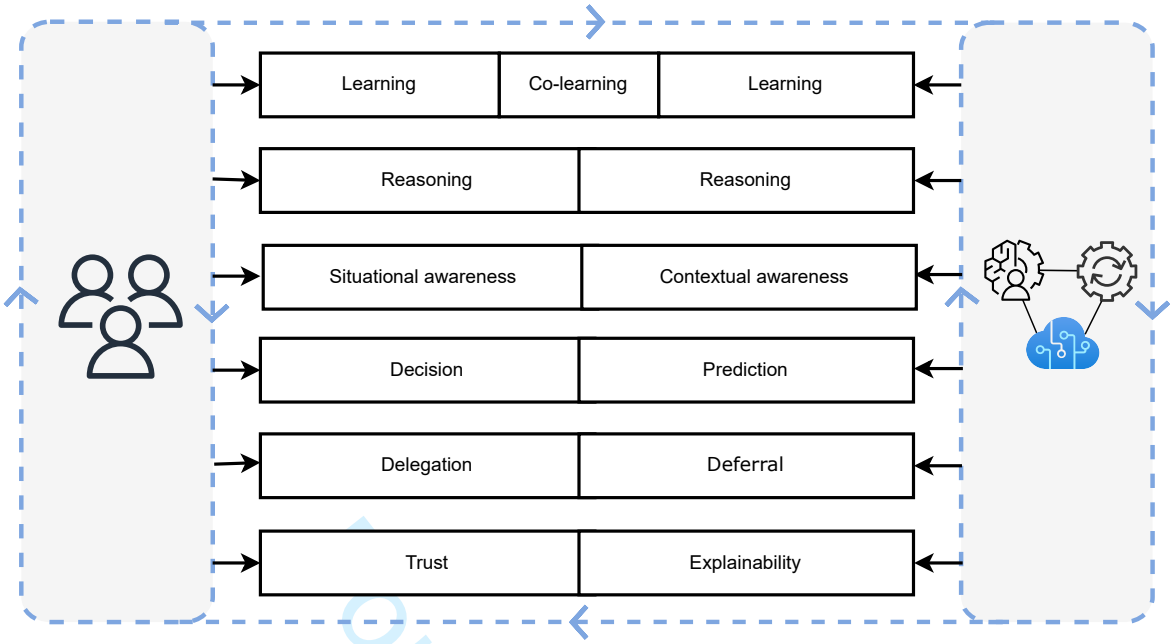

Fig. 2. Key competencies for effective human-AI collaboration. Human competencies include learning, reasoning, situational awareness, decision-making, delegation, and trust. AI competencies mirror and complement these with learning, reasoning, contextual awareness, prediction, deferral, and explainability.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

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 a 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. [61], who compare unidirectional and bidirectional LLMs in biological and chemical domains. Bidirectional models (e.g., BERT [62]) excel at encoding sequences for classification tasks, while unidirectional models (e.g., GPT [63]) 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 GeneAgent, 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 source to determine whether the answer is factually accurate compared to 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 BKGAgent, 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.

Method	Number of agents	Data	Task and goal	Team structure	Tool use	Explainability
ChatNT[64]	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.	NA	NA	NA
DRAGON-AI[65]	No agent	Structured data from existing ontologies and unstructured textual data from sources like GitHub issues	Generate ontological terms.	NA	NA	NA
GeneGPT[66]	No agent	Text data	Answer genomics-related questions by directly generating API request URLs to access and retrieve relevant biomedical information.	NA	NA	NA
Phenomics Assistant[67]	Single AI agent	Monarch[a] knowledge graph	Enhance accessibility to complex genomic information by enabling natural language querying of the Monarch knowledge graph.	NA	Monarch Initiative API	The explainability of AI-generated answers by grounding them in data retrieved from the Monarch KG.
VarChat[68]	Single AI agent	Scientific literature and human genomic variants	Support genetic professionals by providing concise summaries of scientific literature related to specific genomic variants.	NA	Query genomic databases; find and summarize the fragmented scientific literature	Informing users about the sources of its responses.
ChatGSE/ biochatter[69]	Two AI agents	Knowledge graph and scientific articles	Answer user's questions using context from knowledge graphs and scientific papers; demonstrate the usability in cell type annotation task.	Sequential	Information retrieval from knowledge graphs and the literature	Fact-checked and supplemented with context-specific information from documented sources.
BioDiscoveryAgent[70]	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.	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[71]	Two AI agents	GO, Molecular Signature Database (MSigDB[c]), and a proteomics analysis system (NeST[d])	Generate biological process names for gene sets.	Sequential	Call web APIs that connect to biological databases	Providing verification reports that detail the evidence supporting or refuting each generated name.
BRAD[72]	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.
BKGAgent[73]	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[74]	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[75]	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.
[76]	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[77]	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[78]	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.

Table 1. AI assistants in biological and biomedical domains. 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.

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 Machine Learning 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.* [76] 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 [67–69]. The most recent work, Virtual Lab [78], demonstrates the impact of human-AI collaboration 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 [79]. Some AI systems incorporate memory management to continuously track user interactions and dynamically recalibrate the agent's actions based on intermediate results and feedback [67–69, 72]. 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 [78, 80]. It is essential to keep humans in the loop to enhance system performance and reliability [18, 78].

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 (Figure 3C-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 Figure 3B, each assigned distinct roles, as detailed in Figure 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 (Figure 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 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 (Figure 3D), each assigned distinct attributes, including role, perception, and actions (tool use). These attributes enable agents to be optimized for specific domains or functions [81]. 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, 34, 82]. 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 [83].

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.[84], 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 examples

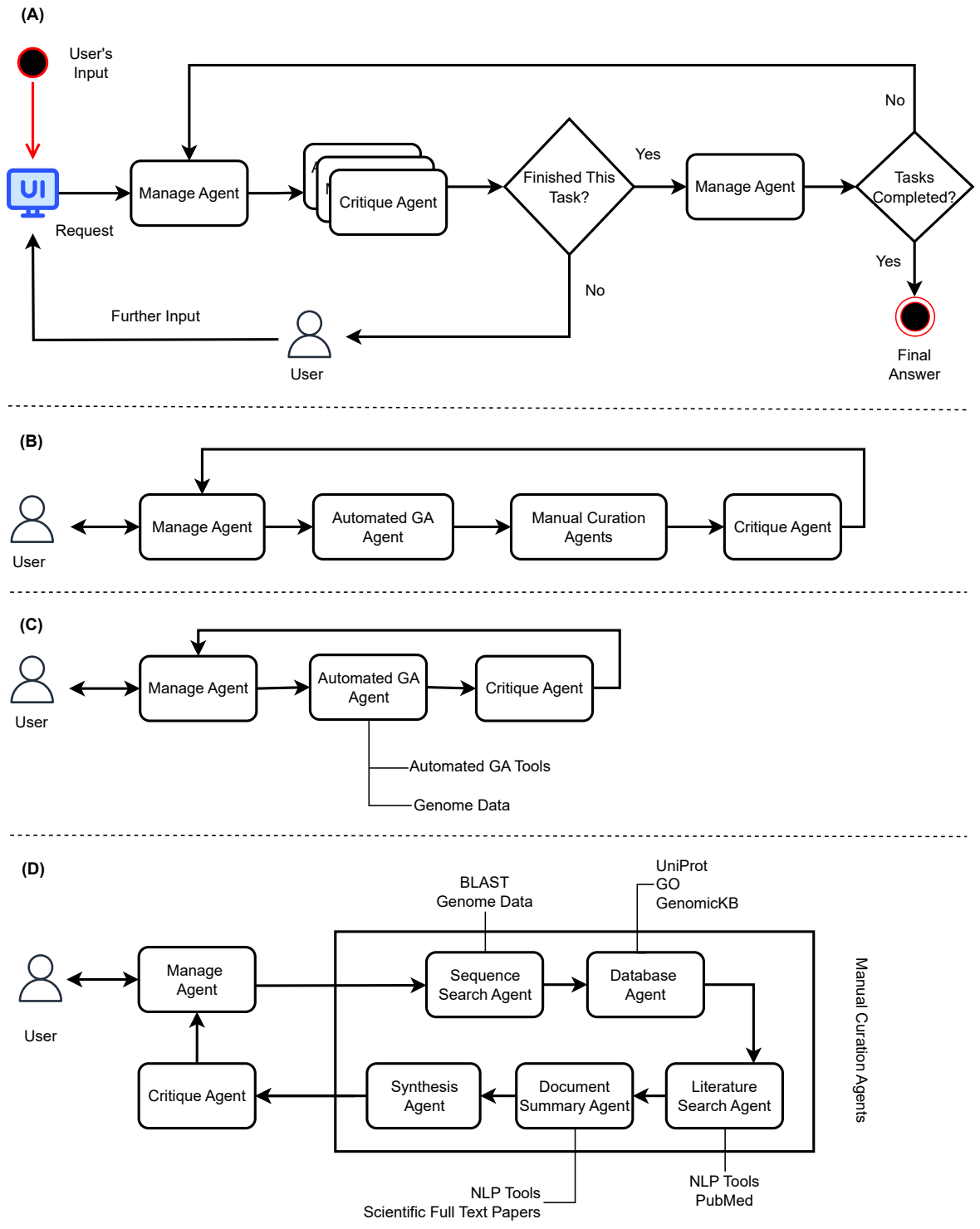

Fig. 3. (A) Overall multi-agent system design for human-AI collaborative genome annotation. Users submit a genome annotation query through an interactive user interface (UI). The UI requests the manager agent to analyze the task, decompose it into subtasks, and assign them to appropriate agents. While 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.

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 genome annotation 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 [85]. 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 [86]. 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 [87].

Developing novel ML/AI methods for enhancing human-AI collaboration

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 [88]. 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 (RLHF) [81].

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 [89]. MemGPT [90] 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 [91].

Requiring multi-dimensional evaluation methods to assess the HAICoGA workflow

Traditional GA evaluation metrics, such as coverage, precision, and accuracy, remain fundamental for assessing annotation quality [3, 92]. 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 [93]. 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 [94].

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 [95]. 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 [94].

Adopting multi-dimensional evaluation frameworks, such as those emphasizing symbiotic HAIC modes, can provide holistic insights [94]. 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 human-AI collaboration

To investigate the challenges and opportunities in CUIs, we developed a chatbot prototype for curating information in gene functional annotation [83]. 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 [96].

Initial testing of the prototype suggests that LLM agents have the potential to serve as valuable tools for collaborative genome annotation (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 [96].

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-centred 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 [97].

To mitigate these risks, a triadic framework involving human regulation, agent alignment, agent regulation and environmental feedback has been proposed [97]. 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 [98]. 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 [99], 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 [98].

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 genome annotation 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 [100, 101], use external tools [102], and answer user questions based on specific contexts [103]. A common architectural foundation for LLMs is the *Transformer* [104], 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 [62]), decoder-only models (e.g., GPT [63]), and encoder-decoder hybrids (e.g., T5 [105]). 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.

Competing interests

No competing interest is declared.

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 a example of HAICoGA framework and future research directions in realize this framework.

References

1. Gökcen Eraslan, Žiga Avsec, Julien Gagneur, and Fabian J Theis. Deep learning: new computational modelling techniques for genomics. Nature Reviews Genetics, 20(7):389–403, 2019.
2. James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. Nature genetics, 51(1):12–18, 2019.
3. Elizabeth H Mahood, Lars H Kruse, and Gaurav D Moghe. Machine learning: a powerful tool for gene function prediction in plants. Applications in Plant Sciences, 8(7):e11376, 2020.
4. Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. Journal of molecular biology, 215(3):403–410, 1990.
5. Mario Stanke, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern. Augustus: ab initio prediction of alternative transcripts. Nucleic acids research, 34(suppl_2):W435–W439, 2006.
6. Mark Yandell and Daniel Ence. A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics, 13(5):329–342, 2012.
7. Ramana Madupu, Lauren M Brinkac, Jennifer Harrow, Laurens G Wilming, Ulrike Böhme, Philippe Lamesch, and Linda I Hannick. Meeting report: a workshop on best practices in genome annotation. Database, 2010:baq001, 2010.
8. Tatiana Tatusova, Michael DiCuccio, Azat Badretdin, Vyacheslav Chetvernin, Eric P Nawrocki, Leonid Zaslavsky, Alexandre Lomsadze, Kim D Pruitt, Mark Borodovsky, and James Ostell. Ncbi prokaryotic genome annotation pipeline. Nucleic acids research, 44(14):6614–6624, 2016.
9. Daniel R Zerbino, Adam Frankish, and Paul Flicek. Progress, challenges, and surprises in annotating the human genome. Annual review of genomics and human genetics, 21:55, 2020.
10. Girum Fitihamlak Ejigu and Jaehee Jung. Review on the computational genome annotation of sequences obtained by next-generation sequencing. Biology, 9(9):295, 2020.
11. Marcela K Tello-Ruiz, Cristina F Marco, Fei-Man Hsu, Rajdeep S Khangura, Pengfei Qiao, Sirjan Sapkota, Michelle C Stitzer, Rachael Wasikowski, Hao Wu, Junpeng Zhan, et al. Double triage to identify poorly annotated genes in maize: The missing link in community curation. PLoS One, 14(10):e0224086, 2019.
12. Eduardo Lee, Gregg A. Helt, Justin T. Reese, Monica C. Munoz-Torres, Chris P. Childers, Robert M. Buels, Lincoln Stein, Ian H. Holmes, Christine G. Elsik, and Suzanna E. Lewis. Web apollo: A web-based genomic annotation editing platform. Genome Biology, 14:1–13, 8 2013.
13. Brian J Haas, Jennifer R Wortman, Catherine M Ronning, Linda I Hannick, Roger K Smith, Rama Maiti, Agnes P Chan, Chunhui Yu, Maryam Farzad, Dongying Wu, et al. Complete reannotation of the arabidopsis genome: methods, tools, protocols and the final release. BMC biology, 3(1):1–19, 2005.
14. Oisin Mac Aodha, Vassilios Stathopoulos, Gabriel J Brostow, Michael Terry, Mark Girolami, and Kate E Jones. Putting the scientist in the loop–accelerating scientific progress with interactive machine learning. In 2014 22nd International Conference on Pattern Recognition, pages 9–17. IEEE, 2014.
15. Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. arXiv preprint arXiv:2404.02831, 2024.

16. Douwe van der Wal, Iny Jhun, Israa Laklouk, Jeff Nirschl, Lara Richer, Rebecca Rojansky, Talent Theparee, Joshua Wheeler, Jörg Sander, Felix Feng, et al. Biological data annotation via a human-augmenting ai-based labeling system. NPJ Digital Medicine, 4(1):145, 2021.

17. Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. Human–computer collaboration for skin cancer recognition. Nature Medicine, 26(8):1229–1234, 2020.

18. Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. Nature Machine Intelligence, pages 1–11, 2024.

19. H James Wilson and Paul R Daugherty. Collaborative intelligence: Humans and ai are joining forces. Harvard Business Review, 96(4):114–123, 2018.

20. Emma Schleiger, Claire Mason, Claire Naughtin, Andrew Reeson, and Cecile Paris. Collaborative intelligence: A scoping review of current applications. Applied Artificial Intelligence, 38(1):2327890, 2024.

21. Ken Goldberg. Robots and the return to collaborative intelligence. Nature Machine Intelligence, 1(1):2–4, 2019.

22. Ming-Hui Huang and Roland T Rust. A framework for collaborative artificial intelligence in marketing. Journal of Retailing, 98(2):209–223, 2022.

23. Jennifer L Reed, Iman Famili, Ines Thiele, and Bernhard O Palsson. Towards multidimensional genome annotation. Nature Reviews Genetics, 7(2):130–141, 2006.

24. Nikos C Kyrpides. Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. Nature biotechnology, 27(7):627–632, 2009.

25. Melissa Hall, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock. A systematic study of bias amplification. arXiv preprint arXiv:2201.11706, 2022.

26. Chia-Yi Cheng, Vivek Krishnakumar, Agnes P Chan, Françoise Thibaud-Nissen, Seth Schobel, and Christopher D Town. Araport11: a complete reannotation of the arabidopsis thaliana reference genome. The Plant Journal, 89(4):789–804, 2017.

27. Philippe Lamesch, Tanya Z Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, Kate Dreher, Debbie L Alexander, Margarita Garcia-Hernandez, et al. The arabidopsis information resource (tair): improved gene annotation and new tools. Nucleic acids research, 40(D1):D1202–D1210, 2012.

28. Suzanna E Lewis, SMJ Searle, Nomi Harris, Mark Gibson, V Iyer, J Richter, C Wiel, L Bayraktaroglu, E Birney, MA Crosby, et al. Apollo: a sequence annotation editor. Genome biology, 3:1–14, 2002.

29. Nathan A Dunn, Deepak R Unni, Colin Diesh, Monica Munoz-Torres, Nomi L Harris, Eric Yao, Helena Rasche, Ian H Holmes, Christine G Elsik, and Suzanna E Lewis. Apollo: democratizing genome annotation. PLoS computational biology, 15(2):e1006790, 2019.

30. Robert Buels, Eric Yao, Colin M Diesh, Richard D Hayes, Monica Munoz-Torres, Gregg Helt, David M Goodstein, Christine G Elsik, Suzanna E Lewis, Lincoln Stein, et al. Jbrowse: a dynamic web platform for genome visualization and analysis. Genome biology, 17(1):1–12, 2016.

31. W James Kent. Blat—the blast-like alignment tool. Genome research, 12(4):656–664, 2002.

32. Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. Pubtator central: automated concept annotation for biomedical full text articles. Nucleic acids research, 47(W1):W587–W593, 2019.

33. Yuqing Mao, Kimberly Van Auken, Donghui Li, Cecilia N Arighi, Peter McQuilton, G Thomas Hayman, Susan Tweedie, Mary L Schaeffer, Stanley JF Laulederkind, Shur-Jen Wang, et al. Overview of the gene ontology task at biocreative iv. Database, 2014:bau086, 2014.

34. Harold J Drabkin, Judith A Blake, and Mouse Genome Informatics Database. Manual gene ontology annotation workflow at the mouse genome informatics database. Database, 2012:bas045, 2012.

35. Mohan Baruwal Chhetri, Shahroz Tariq, Ronal Singh, Fatemeh Jalalvand, Cecile Paris, and Surya Nepal. Towards human-ai teaming to mitigate alert fatigue in security operations centres. ACM Transactions on Internet Technology, 24(3):1–22, 2024.

36. Maria Hartikainen, Guna Spurava, and Kaisa Väänänen. Human-ai collaboration in smart manufacturing: Key concepts and framework for design. HHAI 2024: Hybrid Human AI Systems for the Social Good, pages 162–172, 2024.

37. Antoni Viros-i Martin and Daniel Selva. A framework to study human-ai collaborative design space exploration. In International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, volume 85420, page V006T06A052. American Society of Mechanical Engineers, 2021.

38. Alpana Dubey, Kumar Abhinav, Sakshi Jain, Veenu Arora, and Asha Puttaveerana. 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, pages 1–9, 2020.

39. Dominik Dellermann, Adrian Calma, Nikolaus Lipusch, Thorsten Weber, Sascha Weigel, and Philipp Ebel. The future of human-ai collaboration: a taxonomy of design knowledge for hybrid intelligence systems. arXiv preprint arXiv:2105.03354, 2021.

40. Nicola Cooper. Cognitive biases. ABC of clinical reasoning, pages 22–26, 2017.

41. Gary Klein. Naturalistic decision making. Human factors, 50(3):456–460, 2008.

42. Daniel Kahneman, Paul Slovic, and Amos Tversky. Judgment under uncertainty: Heuristics and biases. Cambridge university press, 1982.

43. Herbert A Simon. Scientific discovery and the psychology of problem solving. In Models of discovery: And other topics in the methods of science, pages 286–303. Springer, 1977.

44. Axel Abels and Tom Lenaerts. Wisdom from diversity: Bias mitigation through hybrid human-llm crowds. arXiv preprint arXiv:2505.12349, 2025.

45. Kimberly Van Auken, Mary L Schaeffer, Peter McQuilton, Stanley JF Laulederkind, Donghui Li, Shur-Jen Wang, G Thomas Hayman, Susan Tweedie, Cecilia N Arighi, James Done, et al. Bc4go: a full-text corpus for the biocreative iv go task. Database, 2014:bau074, 2014.

46. Eugene V Koonin, Michael Y Galperin, Eugene V Koonin, and Michael Y Galperin. Genome annotation and analysis. Sequence—Evolution—Function: Computational Approaches in Comparative Genomics, pages 193–226, 2003.

47. Cécile Paris and Andrew Reeson. What's the secret to making sure AI does not steal your job? work with it, not against it., pages 177–181. Johns Hopkins University Press, Baltimore, 2024.

48. CJ Mungall, S Misra, BP Berman, J Carlson, E Frise, N Harris, B Marshall, S Shu, JS Kaminker, SE Prochnik, et al. An integrated computational pipeline and database to support whole-genome sequence annotation. Genome Biology, 3(12):1–11, 2002.

49. The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. Nucleic Acids Research, 51(D1):D523–D531, 2023.

50. Gene Ontology Consortium. The gene ontology resource: enriching a gold mine. Nucleic acids research, 49(D1):D325–D334, 2021.

51. Rohan Shawn Sunil, Shan Chun Lim, Manoj Itharajula, and Marek Mutwil. The gene function prediction challenge: Large language models and knowledge graphs to the rescue. Current Opinion in Plant Biology, 82:102665, 2024.

52. Bowen Lou, Tian Lu, TS Raghu, and Yingjie Zhang. Unraveling human-ai teaming: A review and outlook. arXiv preprint arXiv:2504.05755, 2025.

53. Lars-Ola Bligård and Anna-Lisa Osvalder. Ccpe: methodology for a combined evaluation of cognitive and physical ergonomics in the interaction between human and machine. Human Factors and Ergonomics in Manufacturing & Service Industries, 24(6):685–711, 2014.

54. Xueyang Feng, Zhi-Yuan Chen, Yujia Qin, Yankai Lin, Xu Chen, Zhiyuan Liu, and Ji-Rong Wen. Large language model–based human–agent collaboration for complex task solving. arXiv preprint arXiv:2402.12914, 2024.

55. Azad M Madni and Carla C Madni. Architectural framework for exploring adaptive human-machine teaming options in simulated dynamic environments. Systems, 6(4):44, 2018.

56. National Academies of Sciences, Engineering, and Medicine. Human-AI Teaming: State-of-the-Art and Research Needs. The National Academies Press, Washington, DC, 2022.

57. Eduardo Salas, Marissa L Shuffler, Amanda L Thayer, Wendy L Bedwell, and Elizabeth H Lazzara. Understanding and improving teamwork in organizations: A scientifically based practical guide. Human resource management, 54(4):599–622, 2015.

58. Jan B. Schmutz, Neal Outland, Sarah Kerstan, Ekaterina Georganta, and Anne Sophie Ulfert. Ai-teaming: Redefining collaboration in the digital era. Current Opinion in Psychology, page 101837, 2024.

59. Thomas A. O'Neill, Nathan J. McNeese, Amy Barron, and Beau G. Schelble. Human–autonomy teaming: A review and analysis of the empirical literature. Human Factors: The Journal of the Human Factors and Ergonomics Society, 64(5):904–938, 2022.

60. Joseph B Lyons, Katia Sycara, Michael Lewis, and August Capiola. Human–autonomy teaming: Definitions, debates, and directions. Frontiers in psychology, 12:589585, 2021.

61. Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. Scientific large language models: A survey on biological & chemical domains. arXiv preprint arXiv:2401.14656, 2024.

62. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pages 4171–4186, 2019.

63. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

64. Guillaume Richard, Bernardo P de Almeida, Hugo Dalla-Torre, Christopher Blum, Lorenz Hexemer, Priyanka Pandey, Stefan Laurent, Marie P Lopez, Alexander Laterre, Maren Lang, et al. Chatnt: A multimodal conversational agent for dna, rna and protein tasks. bioRxiv, pages 2024–04, 2024.

65. Sabrina Toro, Anna V Anagnostopoulos, Sue Bello, Kai Blumberg, Rhiannon Cameron, Leigh Carmody, Alexander D Diehl, Damion Dooley, William Duncan, Petra Fey, et al. Dynamic retrieval augmented generation of ontologies using artificial intelligence (dragon-ai). arXiv preprint arXiv:2312.10904, 2023.

66. Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. Bioinformatics, 40(2):btae075, 2024.

67. Shawn T O'Neil, Kevin Schaper, Glass Elsarboukh, Justin T Reese, Sierra AT Moxon, Nomi L Harris, Monica C Munoz-Torres, Peter N Robinson, Melissa A Haendel, and Christopher J Mungall. Phenomics assistant: An interface for llm-based biomedical knowledge graph exploration. bioRxiv, pages 2024–01, 2024.

68. Federica De Paoli, Silvia Berardelli, Ivan Limongelli, Ettore Rizzo, and Susanna Zucca. Varchat: the generative ai assistant for the interpretation of human genomic variations. Bioinformatics, 40(4):btae183, 2024.

69. Sebastian Lobentanzer, Shaohong Feng, The BioChatter Consortium, Andreas Maier, Cankun Wang, Jan Baumbach, Nils Krehl, Qin Ma, and Julio Saez-Rodriguez. A platform for the biomedical application of large language models. arXiv preprint arXiv:2305.06488, 2023.

70. Yusuf Roohani, Jian Vora, Qian Huang, Zachary Steinhart, Alexander Marson, Percy Liang, and Jure Leskovec. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments. arXiv preprint arXiv:2405.17631, 2024.

71. Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, and Zhiyong Lu. Geneagent: Self-verification language agent for gene set knowledge discovery using domain databases. arXiv preprint arXiv:2405.16205, 2024.

72. Joshua Pickard, Marc Andrew Choi, Natalie Oliven, Cooper Stansbury, Jillian Cwycyshyn, Nicholas Galioto, Alex Gorodetsky, Alvaro Velasquez, and Indika Rajapakse. Bioinformatics retrieval augmentation data (brad) digital assistant. arXiv preprint arXiv:2409.02864, 2024.

73. Xinna Lin, Siqi Ma, Junjie Shan, Xiaojing Zhang, Shell Xu Hu, Tiannan Guo, Stan Z Li, and Kaicheng Yu. Biokgbench: A knowledge graph checking benchmark of ai agent for biomedical science. arXiv preprint arXiv:2407.00466, 2024.

74. Haoyang Liu and Haohan Wang. Genotex: A benchmark for evaluating llm-based exploration of gene expression data in alignment with bioinformaticians. arXiv preprint arXiv:2406.15341, 2024.

75. Alireza Ghafarollahi and Markus J Buehler. Protagents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. Digital Discovery, 3:1389–1409, 2024.

76. Dennis Bersenev, Ayako Yachie, and Sucheendra K Palaniappan. Replicating a high-impact scientific publication using systems of large language models. bioRxiv, pages 2024–04, 2024.

77. Haoyang Liu, Yijiang Li, Jinglin Jian, Yuxuan Cheng, Jianrong Lu, Shuyi Guo, Jinglei Zhu, Mianchen Zhang, Miantong Zhang, and Haohan Wang. Toward a team of ai-made scientists for scientific discovery from gene expression data. arXiv preprint arXiv:2402.12391, 2024.

78. Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab: Ai agents design new sars-cov-2 nanobodies with experimental validation. bioRxiv, pages 2024–11, 2024.

79. Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In International Conference on Learning Representations (ICLR), 2023.

80. Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. arXiv preprint arXiv:2308.03688, 2023.

81. Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu,

Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey. arXiv preprint arXiv:2309.07864, 2023.

82. Erin McDonnell, Kimchi Strasser, and Adrian Tsang. Manual gene curation and functional annotation. In Fungal Genomics, pages 185–208. Springer, 2018.
83. Xiaomei Li, Alex Whan, Meredith McNeil, Samuel C Andrew, Xiang Dai, Madeline Fechner, Cécile Paris, and Rad Suchecki. Genewhisperer: Enhancing manual genome annotation with large language models. bioRxiv, pages 2025–03, 2025.
84. Helena Kudiabor. Virtual lab powered by'ai scientists' super-charges biomedical research. Nature, 636(8043):532–533, 2024.
85. Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. arXiv preprint arXiv:2502.18864, 2025.
86. Thorsten Händler. Balancing autonomy and alignment: A multi-dimensional taxonomy for autonomous llm-powered multi-agent architectures. arXiv preprint arXiv:2310.03659, 2023.
87. Junda He, Christoph Treude, and David Lo. Llm-based multi-agent systems for software engineering: Vision and the road ahead. arXiv preprint arXiv:2404.04834, 2024.
88. Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232, 2023.
89. Kostas Hatalis, Despina Christou, Joshua Myers, Steven Jones, Keith Lambert, Adam Amos-Binks, Zohreh Dannenhauer, and Dustin Dannenhauer. Memory matters: The need to improve long-term memory in llm-agents. In Proceedings of the AAAI Symposium Series, volume 2, pages 277–280, 2023.
90. Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. Memgpt: Towards llms as operating systems. arXiv preprint arXiv:2310.08560, 2023.
91. Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. arXiv preprint arXiv:2212.10403, 2022.
92. Christos A Ouzounis and Peter D Karp. The past, present and future of genome-wide re-annotation. Genome Biology, 3(2):1–6, 2002.
93. Bogdan M Kirilenko, Chetan Munegowda, Ekaterina Osipova, David Jebb, Virag Sharma, Moritz Blumer, Ariadna E Morales, Alexis-Walid Ahmed, Dimitrios-Georgios Kontopoulos, Leon Hilgers, et al. Integrating gene annotation with orthology inference at scale. Science, 380(6643):eabn3107, 2023.
94. George Fragiadakis, Christos Diou, George Kousiouris, and Mara Nikolaidou. Evaluating human-ai collaboration: A review and methodological framework. arXiv preprint arXiv:2407.19098, 2024.
95. Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Is the most accurate ai the best teammate? optimizing ai for teamwork. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 11405–11414, 2021.
96. Melanie McGrath, Patrick Cooper, Andreas Duenser, Jessica Irons, Xiaomei Li, and Rad Suchecki. 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, pages 1–9, 2025.
97. Xiangru Tang, Qiao Jin, Kunlun Zhu, Tongxin Yuan, Yichi Zhang, Wangchunshu Zhou, Meng Qu, Yilun Zhao, Jian Tang, Zhuosheng Zhang, et al. Prioritizing safeguarding over autonomy: Risks of llm agents for science. arXiv preprint

arXiv:2402.04247, 2024.

98. Kaixuan Huang, Yuanhao Qu, Henry Cousins, William A Johnson, Di Yin, Mihir Shah, Denny Zhou, Russ Altman, Mengdi Wang, and Le Cong. Crispr-gpt: An llm agent for automated design of gene-editing experiments. arXiv preprint arXiv:2404.18021, 2024.
99. Kunlun Zhu, Jiaxun Zhang, Ziheng Qi, Nuoxing Shang, Zijia Liu, Peixuan Han, Yue Su, Haofei Yu, and Jiaxuan You. Safescientist: Toward risk-aware scientific discoveries by llm agents. arXiv preprint arXiv:2505.23559, 2025.
100. Hao Wen, Yuanchun Li, Guohong Liu, Shanhui Zhao, Tao Yu, Toby Jia-Jun Li, Shiqi Jiang, Yunhao Liu, Yaqin Zhang, and Yunxin Liu. Empowering llm to use smartphone for intelligent task automation. arXiv preprint arXiv:2308.15272, 2023.
101. Ishika Singh, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg. Progprompt: Generating situated robot task plans using large language models. In 2023 IEEE International Conference on Robotics and Automation (ICRA), pages 11523–11530. IEEE, 2023.
102. Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761, 2023.
103. Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901, 2020.
104. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.
105. Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140):1–67, 2020.

Xiaomei Li is a postdoctoral fellow at CSIRO, specializing in the development of frameworks and methodologies that enhance human-AI collaborative intelligence within the domain of Genomics and Bioinformatics.

Alex Whan is a research scientist at CSIRO, where he focuses on the development of approaches to store and integrate experimental data to improve analysis and insight from biological systems.

Meredith McNeil is a team leader at CSIRO, where her focus is on applying advances in genomics, phenomics, and computational biology to assist in the development of improved crop varieties for Australian agriculture.

David Starns is a genome curation and annotation scientist for the VEuPathDB database resource based at the University of Liverpool. David focuses on parasite, fungal and vector Eukaryote genomes and facilitates scientific community engagement.

Jessica Irons is a behavioural scientist in CSIRO's Data61, researching the development of collaborative human-AI workflows, drawing on principles from human factors and cognitive psychology.

Samuel C. Andrew is a research scientist at CSIRO, exploring how transcriptomic responses to stress and functional traits can be used to understand the thermal tolerance capacity of crops and wild plants.

Rad Suchecki spearheaded the efforts to develop workflows for human-AI collaborative discovery in genomics and bioinformatics during his time as a research scientist and the Genomics and AI team leader at CSIRO Ag & Food. He has since joined Alkahest Inc., as a Sr Computational Biologist working with healthcare and omics data.

Supplementary Notes of 'A Conceptual Framework for Human-AI Collaborative Genome Annotation'

Supplementary Note 1: Automated genome annotation

Evolution of automated genome annotation

Automated genome annotation usually consists of several steps, including sequence alignment, masking repeat sequences, identifying genomic elements, predicting their functions, and performing quality control (McEntyre and Ostell, 2002). Numerous computational methods have been developed for each step of the genome annotation process. A rough chronology can be outlined based on the emergence and prominence of various methods.

In the 1990s and 2000s, rule-based and heuristic methods were the most common approaches for genome annotation. These methods relied on predefined rules or heuristics to identify genes in genome sequences. Tools like BLAST (Altschul et al., 1990) became essential for similarity-based annotations, while gene prediction software such as Glimmer (Delcher et al., 1999) for prokaryotes and AUGUSTUS (Stanke et al., 2003) for eukaryotes utilized statistical models like hidden Markov models (HMMs). These methods are commonly referred to as homology-based (e.g., BLAST) and *ab initio* (e.g., Glimmer and AUGUSTUS) approaches in the literature. Homology-based methods compare the target genome sequence to a database of known genes, which serve as templates. If the target sequence is similar to a known gene, it is likely that the sequence also encodes a gene with similar functions. *Ab initio* methods, on the other hand, detect protein-coding genes by identifying conserved features within the target genome. These conserved features include statistical properties of protein-coding sequences and regulatory signals surrounding protein-coding genes (Tiwari et al., 1997).

In the 2000s and 2010s, integrated and ensemble methods became increasingly popular. These approaches combined multiple sources of evidence to improve prediction accuracy. Tools like MAKER (Holt and Yandell, 2011) integrated *ab initio* gene predictions, protein homology, and transcript evidence. Meanwhile, ensemble methods such as EVidenceModeler (Haas et al., 2008) combined outputs from multiple annotation tools to derive consensus annotations.

In the 2010s and 2020s, ML and deep learning methods have been used to improve the accuracy of genome annotation. Traditional ML techniques, including support vector machines (SVMs) and random forest (RFs), began to be used for various tasks in GA (Mahood et al., 2020). Deep learning methods, such as convolutional neural networks

(CNNs), have been employed for predicting gene structures, regulatory elements, and other genomic features (Mahood et al., 2020; Sapoval et al., 2022; Zou et al., 2019). Well-known examples include AlphaFold3 (Abramson et al., 2024) for biomolecular structure prediction and SpliceAI (Jaganathan et al., 2019) for splice site prediction.

The 2020s have marked a revolutionary era in genome annotation with the rise of Generative AI and advanced machine learning technologies. One prominent example is AlphaFold3 (Abramson et al., 2024), which continues to redefine biomolecular structure prediction with unprecedented accuracy and scalability. Generative AI has further expanded its capabilities with diffusion models that can design proteins with desired properties (Watson et al., 2023). These advancements have significant implications for understanding complex biological processes and engineering novel biomolecules. Large Language Models (LLMs), such as ChatGPT, have gained global attention for their ability to process and generate human language. Their applications extend beyond language processing to the analysis of sequential biological data, including DNA, proteins, and gene expression. In the life sciences, LLMs have been trained on diverse datasets encompassing natural language, molecular sequences, protein structures, and genomic information (Zhang et al., 2024).

New methods are constantly being developed, and existing methods are continually refined in response to growing datasets, our ever-expanding understanding of genomics, and emerging biotechnologies. For example, long-read sequencing techniques, such as those offered by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), have dramatically improved genome assembly (Logsdon et al., 2020). New computational methods are required to process long-read sequencing data (Amarasinghe et al., 2020) or integrate short- and long-read sequencing data (Olson et al., 2022).

Automated genome annotation pipeline

In practice, genome annotation is performed using a genome annotation pipeline, which consists of a series of computational tools. One or more specialized computational tools are employed at each step of the process. These multiple computationally intensive steps are executed in a predefined order, handling various data sources such as aligned sequences, RNA-seq data, and protein information (Holt and Yandell, 2011; McEntyre and Ostell, 2002; Fiddes et al., 2018; Banerjee et al., 2021; Solovyev et al., 2006). Many biological analyses rely on pipelines that have been established and validated over many years. The adoption of new methodologies requires rigorous validation to ensure that new pipelines are both accessible and reliable. While deep learning and advanced AI techniques have seen significant advances and applications in many fields, their full-scale adoption in genome annotation may still be ahead of us.

Challenges in automated genome annotation

Lack of gold-standard data

Since computational tools, particularly ML/AI models, learn from data, the quality of input data directly affects the effectiveness and reliability of the resulting models. However, genome databases, especially those for non-model species, often contain incomplete or inaccurate data. Even in well-studied model species such as Arabidopsis thaliana, a substantial proportion of genes (36%) still lack molecular function or biological process

annotations (TAIR database, as of July 2016) (Bolger et al., 2018). The challenge is even greater for non-model species, where annotations are primarily inferred from model organisms without species-specific experimental validation.

Additionally, public genome databases have been found to contain substantial errors (Bolger et al., 2018; Schnoes et al., 2009). Such errors include, but are not limited to, incorrect gene boundaries, inaccurate exon-intron structures, misidentification of pseudogene regions as genes, and improper functional assignments. Errors can arise at various stages of genome analysis (Bolger et al., 2018; Müller et al., 2003). One of the primary causes of annotation errors is contamination within assemblies, which has been detected in nearly all major databases (Bensch et al., 2021; Bolger et al., 2018). These errors propagate through downstream analyses (Kyrpides, 2009) and are amplified by other ML/AI models (Hall et al., 2022).

Generalizing a model to a new species

The scarcity of diverse datasets significantly hampers a models' ability to generalize and perform effectively across a wide range of genomic contexts. This limitation not only restricts models' training but also undermines their capacity to discover novel patterns and make accurate predictions in less-studied species.

For example, homology-based models transfer the annotations of homologous genes to target genes based on sequence or structural similarity (Korf et al., 2001; Keilwagen et al., 2016; van Baren and Brent, 2006). However, these models struggle to recognize novel genomic elements that lack homologous sequences in existing databases, such as species-specific genes and newly sequenced genomes with long evolutionary distances from reference genomes. The performance of homologous gene annotation depends on the completeness and quality of genome assemblies and the accuracy of reference genome annotations.

Ab initio models are often trained on a specific species (Lomsadze et al., 2005; Burge and Karlin, 1997; Salamov and Solovyev, 2000). They optimize performance by leveraging available signals in the training data, which may be specific to that dataset. As a result, their ability to generalize to new species is impaired.

Integrating multi-omics data.

An increasing number of innovative approaches, including integrated methods (Holt and Yandell, 2011) and hybrid methods (Bruna et al., 2020; Stanke et al., 2006; Solovyev et al., 2006), combine multiple omics data to achieve more accurate and comprehensive genome annotation. However, these methods are often limited by incomplete evidence or insufficient training data, particularly in non-model species, where the quantity and quality of available data for each omic may vary. Additionally, integrating multi-omics data into ML/AI models remains challenging due to the high-dimensional and heterogeneous nature of such data (Barua et al., 2023).

Interpreting ML/AI models.

ML/AI, particularly deep learning algorithms, have recently garnered increasing interest in various genome annotation tasks (Eraslan et al., 2019; Zou et al., 2019). Deep learning methods can significantly improve genome annotation accuracy (Abramson et al., 2024), but their results are more challenging to interpret than those of traditional ML/AI ap-

proaches (Talukder et al., 2021). This complexity obscures the understanding of deep learning models' predictive mechanisms and makes pinpointing potential inaccuracies in the annotation process more difficult.

Supplementary Note 2: Manual curation

Overview of manual curation

Manual genome curation can be categorized into three primary approaches: *museum*, *cottage industry* and *jamboree* (Stein, 2001). The museum approach relies on a centralized team of experts responsible for curating the entire genome, usually for well-studied model species. In the cottage industry approach, multiple independent curators contribute to different regions of the genome. The jamboree approach brings together a large group of participants to collaborate on a genome annotation project. While this approach can expedite the process and reduce costs, it requires robust quality control mechanisms to ensure consistency. Community-based curation (Rödelsperger et al., 2019) and crowdsourcing biocuration (Ramsey et al., 2021) are examples of the jamboree approach.

The manual curation process may vary depending on the specific project. However, some common steps can be followed in manual curation, as shown in Figure 1.

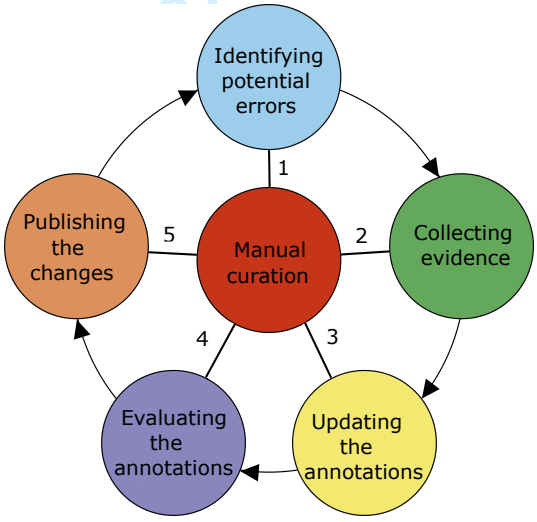

Figure 1: Overview of manual curation.

Step 1: Identifying potential errors in genome annotation Potential errors may occur in features such as assembly gap regions, non-canonical splice sites, multiple alternative splicing forms, trans-spliced genes, and putative gene functions lacking sufficient evidence. Curators can compare genome annotations against external evidence to detect regions that deviate from the expected structure or function of known genes. They can also identify discrepancies where predicted gene models are annotated differently by various pipelines (McDonnell et al., 2018).

Step 2: Collecting evidence to enhance genome annotation To improve the understanding of genomic elements, curators can gather diverse lines of evidence. For instance, experimental data, such as RNA and protein sequences, are useful for confirming the presence or absence of transcripts and determining the exon-intron structure of genes

(Cheng et al., 2017). Furthermore, public databases such as GenBank (Sayers et al., 2020), RefSeq (O'Leary et al., 2016), UniProt (The UniProt Consortium, 2023), InterPro (Paysan-Lafosse et al., 2023), Pfam (Mistry et al., 2021), GO (Gene Ontology Consortium, 2021), and KEGG (Kanehisa and Goto, 2000) provide invaluable information on nucleotide and protein sequences, as well as various functional annotations. Finally, extensive research literature can offer additional experimental gene annotations and sequences that may not be available in public databases.

Step 3: Updating the annotations based on evidence Once evidence is collected, curators can update genome annotations accordingly. For example, gene boundaries can be refined by reviewing RNA or protein sequence alignments. Multiple alternative splicing forms can be identified using RNA-Seq data from various conditions and tissues. Untranslated regions (UTRs) missing from automated GA can be recovered using short peptide data (Madupu et al., 2010). Pseudogenes can be detected by identifying incomplete protein domains in predicted gene models (McDonnell et al., 2018). Such manual curation helps reduce error rates in genome annotation.

Step 4: Evaluating the updated annotations To ensure annotation accuracy, curators should identify any inconsistencies (McDonnell et al., 2018). For example, they can search for homologous genes and assess their functional relationships. Genes with high similarity, belonging to the same gene families or a particular group, should be assigned the same functional annotation. While not an absolute rule, this guideline helps curators maintain consistency in annotations.

Step 5: Publishing the updated annotations and corresponding evidence to public databases Publishing updated annotations in public databases is crucial for enriching scientific knowledge and advancing research (Ramsey et al., 2021). While large-scale projects typically share their updated annotations, smaller projects may focus on identifying genes of interest without immediately submitting their changes to public databases.

Support tools and systems

Online platforms such as GeneDB (Manske et al., 2019) and other Wiki-based systems (listed at https://ngdc.cncb.ac.cn/sciencewikis/index.php/Biological_Wikis) allow registered users to collaboratively create and edit genome annotation pages in real time (Ramsey et al., 2021). This approach decentralizes the curation process and leverages the collective expertise of the global community. However, a key concern with Wiki-based systems is that open editing may introduce errors and biases.

In contrast, standalone and web-based curation systems can be used within more defined teams and projects. For example, the AceDB annotation editor and its variants have been used for C. elegans, the human genome sequence, and the Berkeley Drosophila Genome Project (Consortium, 1999; Frankish and Harrow, 2014). Artemis and the Artemis Comparison Tool (ACT) were primarily designed for reviewing smaller prokaryotic or eukaryotic genomes (Rutherford et al., 2000) but have since been extended to more complex genomes (Carver et al., 2008, 2012). Neomorphic's Annotation Station gene editor was used in The Institute for Genomic Research (TIGR) program for the re-annotation of the Arabidopsis genome (Haas et al., 2005). yrGATE has been used to correct exon-intron structures of genes in several plant-specific databases (Wilkerson et al., 2006). Manual Annotation Studio (MAS) was developed to improve the efficiency of manual functional

annotation of prokaryotic and viral genomes (Lueder et al., 2021). DNA Master has been used for bacteriophage genome annotation (Salisbury and Tsourkas, 2019). MaGe has been used to refine the automatic prediction of gene product functions in bacterial genomes (Vallenet et al., 2006). PeerGAD was a peer-review-based, community-centric web application for viewing and annotating prokaryotic genome sequences (D'Ascenzo et al., 2004). Manatee, developed by The Institute for Genomic Research (TIGR), has been widely used for microbial genome annotation (Haas et al., 2005).

Recently, Apollo has gained widespread adoption within the genome annotation (GA) community (Dunn et al., 2019). Many GA projects integrate this editing tool and use it to annotate various genomes, including GadFly (Mungall et al., 2002), Galaxy (Ramsey et al., 2020) (e.g., G-OnRamp (Liu et al., 2019)), GeneSAS (Humann et al., 2019), DNA Subway (Hilgert et al., 2014), Bovine [Childers et al. (2011); triant2020using], VectorBase (Giraldo-Calderón et al., 2022), VEuPathDB (Amos et al., 2022), and the i5k workspace (Poelchau et al., 2015). Apollo also plays a crucial role in community-driven GA projects, such as the "Genome Decoders" initiative by the Sanger Institute and WormBase, where school students collaborate to annotate the human whipworm genome (Dunn et al., 2019).

These systems provide a platform for aggregating information from diverse databases and curators. While they integrate multiple tools designed for specific tasks, they may still require external tools to support additional functions, particularly for curating information from scientific literature. Scientific literature is invaluable for identifying genes extensively studied in wet-lab experiments, where gene labels are considered the gold standard for gene functions.

Text mining tools have been employed to accelerate two key processes in manual curation: literature searching and information retrieval (Drabkin et al., 2012).

Literature searching tools, such as Textpresso (Müller et al., 2018) and PubSearch (Yoo et al., 2006), identify the most relevant publications related to genes of interest. Users can search the literature using keywords, including gene names, article metadata, and ontology terms. These tools generate indexes that link keywords to articles, enabling curators to efficiently review target genes and their associated publications.

Information retrieval tools extract key concepts for users to review, such as PubTator (Wei et al., 2019), NCBOAnnotator (Tchechmedjiev et al., 2018), Canto (Rutherford et al., 2014), and OntoMate (Liu et al., 2015). Additionally, various natural language processing (NLP) methods can be incorporated into these systems for biological entity recognition, entity linking, and relation extraction (Luo et al., 2023; Lee et al., 2020; Li et al., 2019; Fang et al., 2023). For example, biological entity recognition extracts gene names, mutations, and species from the text; entity linking connects these entities to their corresponding entries in a knowledge base or ontology; and relation extraction identifies relationships between entities, such as gene-disease associations, which are crucial for knowledge discovery.

Supplementary Note 3: Humans competencies

Learning. Human curators engage in domain learning in genomics to make accurate annotation judgement. Additionally, through experience on a task, humans build "mental models", mental representations of the key elements of a task (e.g., theoretical concepts,

task steps and actors) and how these components interact (Cannon-Bowers et al., 1993). This includes an understanding of their teammates and how the team works together to achieve the goal. In the HAICoGA context, humans learn from their experiences working with AI to help understand the strengths and limitations of AI and themselves (Andrews et al., 2022).

Reasoning. Human reasoning is characterized by creativity, intuition, and the ability to deal with unstructured information through experience and learning. (Patterson, 2017, Zheng et al. (2017)). Unlike AI systems, human reasoning excels in abstract thinking and dynamic problem-solving. This ability stems from the human brain's intricate neural structure, enabling the integration of past experiences, contextual understanding, and intuitive judgment. Humans can synthesize disparate pieces of information, make decisions in uncertain conditions, and adapt to new situations swiftly, showcasing a level of flexibility and depth that current AI systems struggle to replicate. This inherent complexity and adaptability make human reasoning an invaluable component in collaborative human-AI systems (Zheng et al., 2017).

Situational awareness. Human situational awareness, as it is commonly defined, refers to an agent's (1) perception of the key information or elements in the task, (2) comprehension of their meaning and (3) predictions about how the situation will unfold (Endsley, 1995). Situational awareness develops dynamically during a task as humans engage in "sense-making" (Klein et al., 2006), and it plays a critical role in task performance (Endsley, 1995). When working with AI, situational awareness includes the ability to monitor the actions of AI and execute corresponding responses, which can increase the effectiveness of collaborations (Jiang et al., 2022). For example, humans will know how best to put prompts to a chatbot by investigating information in conversations (Lou et al., 2023).

Decision Humans make decisions based on their judgment, situation, knowledge or a combination of available data. Considering functional annotations of genes by GO, all annotations need to be supported with statements of evidence and source publications (McDonnell et al., 2018). Human curators need to decide the GO term and evidence code for each annotation using their expertise and a variety of data from the peer-reviewed literature. Humans utilize their judgment in decision-making to navigate complex situations. For example, even with curator expertise and supporting evidence, full certainty cannot be achieved. Curators' judgments lead them to make decisions under some uncertainty (with lower confidence) or postpone decision-making until new evidence is available. Moreover, whether to accept an AI's suggestion is often left to individual curators' judgments and is likely not very standardized.

Delegation. Humans can delegate tasks to leverage complementary AI competencies(Pinski et al., 2023). Deciding to delegate to AI is based on, among other things, the human's assessment of the AI's ability and confidence in their own (Fügener et al., 2019). Due to the speed, scalability and quantitative capabilities of AI, humans usually make AI automatically annotate a large scale of the genome, or require AI to provide relevant information about gene annotations by searching databases and literature.

Trust. Ensuring appropriate levels of trust is particularly important to establish successful collaborative relationships between humans and AI. Ideally, the human-AI system should support "calibrated" trust that aligns accurately with the AI's capabilities (Lee and See, 2004). In long-term collaborations, this calibrated trust becomes even more

critical, as it ensures a balanced reliance on AI over time. Trust should be neither too low, causing humans to disengage from the AI, nor too high, leading to overreliance and potential failure to detect errors. Maintaining this balance is essential for sustained collaboration, where the dynamics of trust evolve with ongoing interactions and experiences.

Supplementary Note 4: AI competencies

Learning. AI systems are designed to learn from data and develop their ability to perform tasks using various strategies, such as supervised (Scalzitti et al., 2020), semi-supervised (Jia et al., 2021), and unsupervised learning (Abeel et al., 2008), all of which have already been applied in genome annotation. In HAICoGA, we may leverage additional AI learning strategies, such as reinforcement learning, continuous learning, and active learning, to enhance AI-human collaboration.

Reinforcement learning is a valuable strategy for enhancing human-AI collaboration by enabling AI to learn from and work alongside humans (Navidi and Landry Jr, 2021). It can help AI make decisions and take actions aligned with team goals, adapt to the environment, and improve overall team performance.

Continuous learning refers to an AI system's ability to learn from new data streams in real time without requiring complete retraining (Wang et al., 2022). This capability is crucial for human-AI collaboration, particularly in dynamic environments where conditions and information constantly evolve. Through continuous learning, AI can adapt to new scenarios, update its knowledge, and refine its decision-making processes in response to changing inputs. This flexibility ensures that AI remains relevant and effective over time, enhancing its ability to support humans.

Active learning involves the strategic selection of data samples for human labeling to maximize information gain while minimizing human effort (van der Wal et al., 2021). A potential use case is AI identifying genomic regions with high uncertainty but significant relevance to GA. Humans can then focus their efforts on these informative regions rather than the entire genome.

Co-learning. Co-learning refers to the process in which humans and AI learn from each other through collaborative interactions (van den Bosch et al., 2019). The benefits of human-AI co-learning have been demonstrated in various studies. For example, one study proposed that co-learning fosters mutual understanding, mutual benefits, and mutual growth between humans and AI, ultimately enhancing productivity and creativity (Huang et al., 2019). Another study suggested that co-learning enables humans and AI to discover and understand the task, environment, themselves, and their teammates (Schoonderwoerd et al., 2022). Several methods have been developed to facilitate AI co-learning with humans and achieve optimal performance by integrating various learning strategies (Mozannar and Sontag, 2020; van der Wal et al., 2021).

Reasoning. AI reasoning usually excels in logicality, repeatability, and the efficient processing of structured data through predefined rules and machine learning algorithms (Zheng et al., 2017). This enables AI systems to handle large volumes of data and perform complex calculations at speeds unattainable by humans. In the context of human-AI collaboration, enhancing AI reasoning involves developing systems capable of learning from diverse, dynamic, and unstructured environments. Another critical aspect of improving

AI reasoning is the development of causal models, which allow AI to predict relationships between variables beyond mere correlations (Shipley, 2016; Hill et al., 2016). Causal reasoning helps AI interpret cause-effect relationships, making decision-making processes more transparent and explainable (Zheng et al., 2017). Causal modeling has been instrumental in mitigating selection bias, particularly in genome annotation, by enabling the identification of true causal variants and reducing false positive predictions (Ramstein and Buckler, 2022).

Contextual awareness. AI's contextual awareness refers to its ability to adapt outputs based on the specific task, input data, and surrounding context within a given domain. Context-aware AI systems are designed to dynamically adjust their behavior, defer uncertain cases to human experts, and incorporate user feedback to refine future predictions (Jiang et al., 2023). This capability is essential for ensuring that AI-generated annotations are not only accurate but also biologically meaningful and relevant to the task at hand.

Prediction. AI systems play a critical role in predicting gene structures and functions, while human experts review and refine these predictions to ensure they align with existing evidence and domain knowledge. Beyond prediction, AI also supports tasks such as clustering, information retrieval, processing user queries, and generating context-relevant text, each contributing to different aspects of genome annotation. Additionally, AI system are expected to help flag potential error. Tools like quality metrics and gene tree visualizers exemplify this function by helping to identify misannotations that require human correction (Tello-Ruiz et al., 2019).

Deferral. AI systems are expected to defer certain tasks to humans rather than risk generating incorrect annotations. For example, when encountering challenging cases, such as very short genes, pseudogenes, or tandem gene duplications, AI can defer them to human experts (Madupu et al., 2010). In the current manual curation process, genes lacking GO annotations, those with non-canonical splice sites, or poorly annotated genes are already commonly assigned to human curators (Tello-Ruiz et al., 2019). These are examples of one-off or static deferral strategies. With the advancement of AI methodologies (Mozannar and Sontag, 2020), it is now possible to develop intrinsic deferral mechanisms, enabling AI systems to dynamically defer tasks to humans as part of an ongoing, adaptive collaboration.

Explainability. The explainability of AI refers to its ability to provide a human-understandable rationale for its results. "White-box" models, such as those based on patterns, rules, or decision trees, explicitly explain how they generate predictions and have been widely used in biomedical research. However, "black-box" models, particularly deep learning models, outperform white-box models in complex tasks but lack the transparency needed to produce explainable knowledge (Dey et al., 2022). This can pose challenges in understanding biological mechanisms and building trust in AI-generated conclusions (Ali et al., 2023). These challenges are being addressed through explainable artificial intelligence (XAI) techniques, which enhance the interpretability of black-box models. Such advancements boost confidence in AI-generated outputs across a wide range of biological applications (Zhou et al., 2023; Koh et al., 2024).

Supplementary Note 5: Illustrative use case – Enhancing annotation consistency with an AI-Assisted ontology finder

The following use case scenario illustrates how ChatGPT can assist users in assigning GO terms and Evidence and Conclusion Ontology (ECO) codes based on literature excerpts. The collaboration is presented in a user–ChatGPT 4o (accessed on 12 June 2025) dialogue format to demonstrate the potential of LLMs in supporting ontology-based annotation tasks. The full dialogue can also be accessed at: https://chatgpt.com/share/684a2eac-2400-800b-9036-5a4f9f1cd0fa. Please note that while the final suggestion for the GO term is correct, the associated ECO term appears to be misaligned. The model suggested ECO:0000315, which corresponds to "mutant phenotype evidence used in manual assertion". However, based on the description of direct assay methods in the input text, the appropriate ECO term should be ECO:0000314 ("direct assay evidence used in manual assertion").

User: You are a knowledgeable and reliable AI assistant specializing in biomedical curation. Your primary task is to help human curators identify accurate Gene Ontology (GO) terms and Evidence and Conclusion Ontology (ECO) codes based solely on the provided sentence, paragraph, or experimental description.

- Use only the information available in the input. Do not assume or infer beyond what is stated.
- If the information provided is insufficient to confidently assign a GO or ECO term, politely request clarification or additional context from the human user.
- When suggesting GO or ECO terms, explain your reasoning clearly and concisely, including the exact evidence phrase(s) you used.
- If multiple GO or ECO terms are plausible, present them all and explain the differences.
- Always remain transparent about limitations and defer final judgment to the human curator.

Your role is to enhance consistency, accuracy, and efficiency in ontology-based annotation while ensuring human experts remain in control of final decisions.

ChatGPT 4o: Understood. I'll follow these principles when reviewing experimental descriptions and recommending GO terms and ECO codes. Please share a sentence, paragraph, or description of the experiment you'd like me to analyze. I'll identify candidate GO terms and ECO codes using only the information provided, explain my reasoning clearly, and request clarification if needed.

User: Sentence: "Moreover, the mean cell number of blastocysts was significantly higher in the 40 ng pRLN/ml group (38 $\pm$ 3) compared to others (control: 31 $\pm$ 4 and 20 ng pRLN/ml: 32 $\pm$ 6; $P < 0.05$), which appeared similar.". Can you suggest a GO term and an ECO term related to pRLN based on this sentence?

ChatGPT 4o: Thank you. Based on the sentence:

"Moreover, the mean cell number of blastocysts was significantly higher in the 40 ng pRLN/ml group (38 $\pm$ 3) compared to others (control: 31 $\pm$ 4 and 20 ng pRLN/ml: 32 $\pm$ 6; $P < 0.05$), which appeared similar."

Key evidence phrases: "mean cell number of blastocysts was significantly higher"

"40 ng pRLN/ml group"

Suggests that pRLN (presumably prolactin or relaxin, if pRLN stands for one of these) enhances blastocyst cell proliferation or development

Suggested Gene Ontology (GO) Term(s): GO:0008283 – Cell proliferation Reasoning: The sentence reports a significant increase in cell number in blastocysts treated with 40 ng pRLN/ml. This directly supports a role in promoting cell proliferation.

Used evidence phrase: "mean cell number of blastocysts was significantly higher"

(Alternative, context-dependent): If the focus is more on developmental progression of blastocysts:

GO:0009790 – Embryo development Caution: This term would require more direct evidence showing broader developmental changes, not just cell number.

Suggested Evidence and Conclusion Ontology (ECO) Term: ECO:0000315 – Experimentally inferred by direct assay (IDA) Reasoning: The evidence is derived from a direct measurement of cell number in response to a specific treatment (pRLN), which fits the criteria for IDA.

Used evidence phrase: "mean cell number… significantly higher… compared to others… $P < 0.05$"

Limitations & Clarification Needed: It's unclear what "pRLN" stands for—presumably a peptide or hormone (e.g., prolactin or relaxin). The precise GO annotation might change based on knowing its identity and mechanism.

Also, we don't know if cell type identity or viability was assessed, which could refine annotation.

Summary:

Ontology	Suggested Term	ID	Justification
GO	Cell proliferation	GO:0008283	Based on increased blastocyst cell number
ECO	Inferred from direct assay	ECO:0000315	Based on quantification of cell number in treated vs. control groups

Let me know if more context is available (e.g., full figure legend, experimental method, identity of pRLN), and I can refine the suggestion further.

User:

Thanks, this is Table 4 with details of experimental results.

ChatGPT 4o: Thank you for sharing Figure/Table 4. With this additional context, I can now refine the GO and ECO term suggestions based on the detailed experimental results evaluating relaxin (pRLN) during oocyte maturation (IVM) and embryo culture (IVC).

Key Observations from the Table: Experimental variable: Relaxin (pRLN) concentration during IVM and IVC.

Table 4.

Effects of relaxin added during both oocyte maturation and embryo culture

pRLN (ng/ml) during		Total Zygotes (N)	% of cleaved at Day 2pi (n)	Blastocyst formation at Day 7pi		
IVM	IVC			Total (T)	% (T/N)	Cell number (n)
0	0	324	48 ± 3 (156)[a]	24	8 ± 2[a]	31 ± 3 (24)[a]
40	20	272	33 ± 5 (91)[b]	19	7 ± 2[a]	33 ± 3 (12)[a]
40	40	353	51 ± 8 (172)[a]	33	10 ± 2[a]	43 ± 6 (10)[b]

Open in a new tab

[abc] Different superscripts within the same columns indicate significant difference (P < 0.05; ANOVA). Data are mean values (± SEM) of at least 5 independent replicates.

Figure 2: Table 4 from the scientific paper by Feugang et al. (2011).

Outcomes measured:

Cleavage rate at Day 2 post-insemination

Blastocyst formation rate at Day 7

Blastocyst cell number (proxy for growth/proliferation)

Updated Gene Ontology (GO) Suggestions: GO:0008283 – Cell proliferation

Justification: The mean cell number in blastocysts was significantly higher in the 40 ng/ml pRLN IVM + 40 ng/ml IVC condition (43 ± 6) compared to control (31 ± 3). This indicates increased proliferation in the presence of relaxin.

Evidence phrase: "Cell number (n): 43 ± 6 vs. 31 ± 3"

GO:0009790 – Embryo development

Justification: Relaxin appears to influence not only proliferation but also cleavage rates and blastocyst formation, suggesting an impact on overall developmental progression.

Evidence phrase: "% of cleaved at Day 2pi" and "Blastocyst formation at Day 7pi"

(Optional, mechanistic) If further molecular evidence is available in the full article:

GO:0040018 – Positive regulation of multicellular organism growth May apply if authors link relaxin mechanistically to embryo growth.

Suggested Evidence and Conclusion Ontology (ECO) Term: ECO:0000315 – Experimentally inferred by direct assay (IDA)

Justification: The experiment involved direct measurement of cleavage rates, blastocyst formation, and cell number under different treatment conditions.

Applies to all GO terms above

Summary Table:

Ontology	Term Name	ID	Justification
GO	Cell proliferation	GO:0008283	Increased blastocyst cell number with pRLN
GO	Embryo development	GO:0009790	Effects on cleavage and blastocyst formation
ECO	Inferred from direct assay	ECO:0000315	Direct measurements of cell number and development

Notes for Curator: If molecular mechanisms or gene expression analyses related to relaxin are included elsewhere in the paper, more specific GO terms may apply (e.g., signaling pathways).

The exact GO annotation for relaxin's biological role may also depend on how the authors interpret its function (e.g., hormone activity vs. growth factor activity).

Let me know if you'd like help annotating a specific gene/protein involved in this experiment or more detailed parsing of the full text.

User: Thank you. Table 4 demonstrates that the addition of pRLN to in vitro culture increases the cellularity of blastocysts. Is there a specific Gene Ontology (GO) term that describes this function?

ChatGPT 4o: Yes, there is a specific Gene Ontology (GO) term that describes the function observed in Table 4, where the addition of pRLN to in vitro culture increases the cellularity of blastocysts.

Recommended GO Term: GO:0001832 – Blastocyst growth

Definition: An increase in size of a blastocyst due to expansion of the blastocoelic cavity, cell shape changes, and cell proliferation.

Justification: The observed increase in blastocyst cell number with pRLN treatment reflects enhanced growth through cell proliferation, aligning with this term.

Recommended Evidence Code (ECO): ECO:0000315 – Inferred from Direct Assay (IDA)

Justification: The conclusion is based on direct measurements of blastocyst cell numbers under varying pRLN concentrations, fitting the criteria for IDA.

Summary:

Ontology	Term Name	ID	Justification
GO	Blastocyst growth	GO:0001832	Reflects increased blastocyst size due to cell proliferation
ECO	Inferred from direct assay	ECO:0000315	Based on direct measurement of blastocyst cell numbers

If further details about the specific cell lineages affected (e.g., trophectoderm or inner cell mass) are available, more precise GO terms could be considered. Please provide additional information if available.

References

Abeel, Thomas, Yvan Saeys, Pierre Rouzé, and Yves Van de Peer (2008), "ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles." *Bioinformatics*, 24, i24.

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

Ali, Sajid, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera (2023), "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence." *Information fusion*, 99, 101805.

Altschul, Stephen F, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman (1990), "Basic local alignment search tool." *Journal of molecular biology*, 215, 403–410.

Amarasinghe, Shanika L, Shian Su, Xueyi Dong, Luke Zappia, Matthew E Ritchie, and Quentin Gouil (2020), "Opportunities and challenges in long-read sequencing data analysis." *Genome biology*, 21, 1–16.

Amos, Beatrice, Cristina Aurrecoechea, Matthieu Barba, Ana Barreto, Evelina Y Basenko, Robert Belnap, Ann S Blevins, Ulrike Böhme, John Brestelli, Brian P Brunk, et al. (2022), "Veupathdb: the eukaryotic pathogen, vector and host bioinformatics resource center." *Nucleic Acids Research*, 50, D898–D911.

Andrews, Robert W, J Mason Lilly, Divya Srivastava, and Karen M Feigh (2022), "The role of shared mental models in human-ai teams: a theoretical review." *Theoretical Issues in Ergonomics Science*, 1–47.

Banerjee, Sagnik, Priyanka Bhandary, Margaret Woodhouse, Taner Z Sen, Roger P Wise, and Carson M Andorf (2021), "Finder: an automated software package to annotate eukaryotic genes from rna-seq data and associated protein sequences." *BMC bioinformatics*, 22, 1–26.

Barua, Arnab, Mobyen Uddin Ahmed, and Shahina Begum (2023), "A systematic literature review on multimodal machine learning: Applications, challenges, gaps and future directions." *IEEE Access*, 11, 14804–14831.

Bensch, Staffan, Mizue Inumaru, Yukita Sato, Larisa Lee Cruz, Andrew A Cunningham, Simon J Goodman, Iris I Levin, Patricia G Parker, Patricia Casanueva, Maria-Angeles Hernández, et al. (2021), "Contaminations contaminate common databases." *Molecular ecology resources*, 21, 355–362.

Bolger, Marie E, Borjana Arsova, and Björn Usadel (2018), "Plant genome and transcriptome annotations: from misconceptions to simple solutions." *Briefings in bioinformatics*, 19, 437–449.

Bruna, Tomáš, Alexandre Lomsadze, and Mark Borodovsky (2020), "Genemark-ep+: eukaryotic gene prediction with self-training in the space of genes and proteins." *NAR genomics and bioinformatics*, 2, lqaa026.

Burge, Chris and Samuel Karlin (1997), "Prediction of complete gene structures in human genomic dna." *Journal of molecular biology*, 268, 78–94.

Cannon-Bowers, Janis A, Eduardo Salas, and Sharolyn Converse (1993), "Shared mental models in expert team decision making." *Individual and group decision making: Current issues*, 221–246.

Carver, Tim, Matthew Berriman, Adrian Tivey, Chinmay Patel, Ulrike Böhme, Barclay G Barrell, Julian Parkhill, and Marie-Adèle Rajandream (2008), "Artemis and act: viewing, annotating and comparing sequences stored in a relational database." *Bioinformatics*, 24, 2672–2676.

Carver, Tim, Simon R Harris, Matthew Berriman, Julian Parkhill, and Jacqueline A McQuillan (2012), "Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data." *Bioinformatics*, 28, 464–469.

Cheng, Chia-Yi, Vivek Krishnakumar, Agnes P Chan, Françoise Thibaud-Nissen, Seth Schobel, and Christopher D Town (2017), "Araport11: a complete reannotation of the arabidopsis thaliana reference genome." *The Plant Journal*, 89, 789–804.

Childers, Christopher P, Justin T Reese, Jaideep P Sundaram, Donald C Vile, C Michael Dickens, Kevin L Childs, Hanni Salih, Anna K Bennett, Darren E Hagen, David L Adelson, et al. (2011), "Bovine genome database: integrated tools for genome annotation and discovery." *Nucleic acids research*, 39, D830–D834.

Consortium, FlyBase (1999), "The flybase database of the drosophila genome projects and community literature." *Nucleic Acids Research*, 27, 85–88.

Delcher, Arthur L, Douglas Harmon, Simon Kasif, Owen White, and Steven L Salzberg (1999), "Improved microbial gene identification with glimmer." *Nucleic acids research*, 27, 4636–4641.

Dey, Sanjoy, Prithwish Chakraborty, Bum Chul Kwon, Amit Dhurandhar, Mohamed Ghalwash, Fernando J Suarez Saiz, Kenney Ng, Daby Sow, Kush R Varshney, and Pablo Meyer (2022), "Human-centered explainability for life sciences, healthcare, and medical informatics." *Patterns*, 3, 100493.

Drabkin, Harold J, Judith A Blake, and Mouse Genome Informatics Database (2012), "Manual gene ontology annotation workflow at the mouse genome informatics database." *Database*, 2012, bas045.

Dunn, Nathan A, Deepak R Unni, Colin Diesh, Monica Munoz-Torres, Nomi L Harris, Eric Yao, Helena Rasche, Ian H Holmes, Christine G Elsik, and Suzanna E Lewis (2019), "Apollo: democratizing genome annotation." *PLoS computational biology*, 15, e1006790.

D'Ascenzo, Mark D, Alan Collmer, and Gregory B Martin (2004), "Peergad: a peer-review-based and community-centric web application for viewing and annotating prokaryotic genome sequences." *Nucleic acids research*, 32, 3124–3135.

Endsley, Mica R (1995), "Toward a theory of situation awareness in dynamic systems." *Human factors*, 37, 32–64.

Eraslan, Gökcen, Žiga Avsec, Julien Gagneur, and Fabian J Theis (2019), "Deep learning: new computational modelling techniques for genomics." *Nature Reviews Genetics*, 20, 389–403.

Fang, Li, Qingyu Chen, Chih-Hsuan Wei, Zhiyong Lu, and Kai Wang (2023), "Bioformer: an efficient transformer language model for biomedical text mining." *arXiv preprint arXiv:2302.01588*.

Feugang, Jean M, Jonathan M Greene, Scott T Willard, and Peter L Ryan (2011), "In vitro effects of relaxin on gene expression in porcine cumulus-oocyte complexes and developing embryos." *Reproductive Biology and Endocrinology*, 9, 1–10.

Fiddes, Ian T, Joel Armstrong, Mark Diekhans, Stefanie Nachtweide, Zev N Kronenberg, Jason G Underwood, David Gordon, Dent Earl, Thomas Keane, Evan E Eichler, et al. (2018), "Comparative annotation toolkit (cat)—simultaneous clade and personal genome annotation." *Genome research*, 28, 1029–1038.

Frankish, Adam and Jennifer Harrow (2014), "Gencode pseudogenes." In *Pseudogenes*, 129–155, Springer.

Fügener, Andreas, Jörn Grahl, Alok Gupta, and Wolfgang Ketter (2019), "Cognitive challenges in human-ai collaboration: Investigating the path towards productive delegation." *Forthcoming, Information Systems Research*, 1–39.

Gene Ontology Consortium (2021), "The gene ontology resource: enriching a gold mine." *Nucleic acids research*, 49, D325–D334.

Giraldo-Calderón, Gloria I, Omar S Harb, Sarah A Kelly, Samuel SC Rund, David S Roos, and Mary Ann McDowell (2022), "Vectorbase. org updates: bioinformatic resources for invertebrate vectors of human pathogens and related organisms." *Current opinion in insect science*, 50, 100860.

Haas, Brian J, Steven L Salzberg, Wei Zhu, Mihaela Pertea, Jonathan E Allen, Joshua Orvis, Owen White, C Robin Buell, and Jennifer R Wortman (2008), "Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments." *Genome biology*, 9, 1–22.

Haas, Brian J, Jennifer R Wortman, Catherine M Ronning, Linda I Hannick, Roger K Smith, Rama Maiti, Agnes P Chan, Chunhui Yu, Maryam Farzad, Dongying Wu, et al. (2005), "Complete reannotation of the arabidopsis genome: methods, tools, protocols and the final release." *BMC biology*, 3, 1–19.

Hall, Melissa, Laurens van der Maaten, Laura Gustafson, and Aaron Adcock (2022), "A systematic study of bias amplification." *arXiv preprint arXiv:2201.11706*.

Hilgert, Uwe, Sheldon McKay, Mohammed Khalfan, Jason Williams, Cornel Ghiban, and David Micklos (2014), "Dna subway: making genome analysis egalitarian." In *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment*, 1–3.

Hill, Steven M, Laura M Heiser, Thomas Cokelaer, Michael Unger, Nicole K Nesser, Daniel E Carlin, Yang Zhang, Artem Sokolov, Evan O Paull, Chris K Wong, et al. (2016), "Inferring causal molecular networks: empirical assessment through a community-based effort." *Nature methods*, 13, 310–318.

Holt, Carson and Mark Yandell (2011), "Maker2: an annotation pipeline and genome-database management tool for second-generation genome projects." *BMC bioinformatics*, 12, 1–14.

Huang, Yi-Ching, Yu-Ting Cheng, Lin-Lin Chen, and Jane Yung-jen Hsu (2019), "Human-ai co-learning for data-driven ai." *arXiv preprint arXiv:1910.12544*.

Humann, Jodi L, Taein Lee, Stephen Ficklin, and Dorrie Main (2019), "Structural and functional annotation of eukaryotic genomes with gensas." In *Gene prediction*, 29–51, Springer.

Jaganathan, Kishore, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F McRae, Siavash Fazel Darbandi, David Knowles, Yang I Li, Jack A Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B Schwartz, et al. (2019), "Predicting splicing from primary sequence with deep learning." *Cell*, 176, 535–548.

Jia, Hao, Sung-Joon Park, and Kenta Nakai (2021), "A semi-supervised deep learning approach for predicting the functional effects of genomic non-coding variations." *BMC bioinformatics*, 22, 1–12.

Jiang, Jinglu, Alexander J Karran, Constantinos K Coursaris, Pierre-Majorique Léger, and Joerg Beringer (2022), "A situation awareness perspective on human-ai interaction: Tensions and opportunities." *International Journal of Human–Computer Interaction*, 1–18.

Jiang, Na, Xiaohui Liu, Hefu Liu, Eric Tze Kuan Lim, Chee-Wee Tan, and Jibao Gu (2023), "Beyond ai-powered context-aware services: the role of human–ai collaboration." *Industrial Management & Data Systems*, 123, 2771–2802.

Kanehisa, Minoru and Susumu Goto (2000), "Kegg: kyoto encyclopedia of genes and genomes." *Nucleic acids research*, 28, 27–30.

Keilwagen, Jens, Michael Wenk, Jessica L Erickson, Martin H Schattat, Jan Grau, and Frank Hartung (2016), "Using intron position conservation for homology-based gene prediction." *Nucleic acids research*, 44, e89–e89.

Klein, Gary, Brian Moon, and Robert R Hoffman (2006), "Making sense of sensemaking 1: Alternative perspectives." *IEEE intelligent systems*, 21, 70–73.

Koh, Eugene, Rohan Shawn Sunil, Hilbert Yuen In Lam, and Marek Mutwil (2024),

"Harnessing big data and artificial intelligence to study plant stress." *arXiv preprint arXiv:2404.15776*.

Korf, Ian, Paul Flicek, Daniel Duan, and Michael R Brent (2001), "Integrating genomic homology into gene structure prediction." *Bioinformatics*, 17, S140–S148.

Kyrpides, Nikos C (2009), "Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream." *Nature biotechnology*, 27, 627–632.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang (2020), "Biobert: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics*, 36, 1234–1240.

Lee, John D and Katrina A See (2004), "Trust in automation: Designing for appropriate reliance." *Human factors*, 46, 50–80.

Li, Xiaoya, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li (2019), "A unified mrc framework for named entity recognition." *arXiv preprint arXiv:1910.11476*.

Liu, Weisong, Stanley JF Laulederkind, G Thomas Hayman, Shur-Jen Wang, Rajni Nigam, Jennifer R Smith, Jeff De Pons, Melinda R Dwinell, and Mary Shimoyama (2015), "Ontomate: a text-mining tool aiding curation at the rat genome database." *Database*, 2015, bau129.

Liu, Yating, Luke Sargent, Wilson Leung, Sarah CR Elgin, and Jeremy Goecks (2019), "G-onramp: a galaxy-based platform for collaborative annotation of eukaryotic genomes." *Bioinformatics*, 35, 4422–4423.

Logsdon, Glennis A, Mitchell R Vollger, and Evan E Eichler (2020), "Long-read human genome sequencing and its applications." *Nature Reviews Genetics*, 21, 597–614.

Lomsadze, Alexandre, Vardges Ter-Hovhannisyan, Yury O Chernoff, and Mark Borodovsky (2005), "Gene identification in novel eukaryotic genomes by self-training algorithm." *Nucleic acids research*, 33, 6494–6506.

Lou, Renze, Kai Zhang, and Wenpeng Yin (2023), "Is prompt all you need? no. a comprehensive and broader view of instruction learning." *arXiv preprint arXiv:2303.10475*.

Lueder, Matthew R, Regina Z Cer, Miles Patrick, Logan J Voegtly, Kyle A Long, Gregory K Rice, and Kimberly A Bishop-Lilly (2021), "Manual annotation studio (mas): A collaborative platform for manual functional annotation of viral and microbial genomes." *BMC genomics*, 22, 1–14.

Luo, Ling, Chih-Hsuan Wei, Po-Ting Lai, Robert Leaman, Qingyu Chen, and Zhiyong Lu (2023), "Aioner: all-in-one scheme-based biomedical named entity recognition using deep learning." *Bioinformatics*, 39, btad310.

Madupu, Ramana, Lauren M Brinkac, Jennifer Harrow, Laurens G Wilming, Ulrike Böhme, Philippe Lamesch, and Linda I Hannick (2010), "Meeting report: a workshop on best practices in genome annotation." *Database*, 2010, baq001.

Mahood, Elizabeth H, Lars H Kruse, and Gaurav D Moghe (2020), "Machine learning: a powerful tool for gene function prediction in plants." *Applications in Plant Sciences*, 8, e11376.

Manske, Magnus, Ulrike Böhme, Christoph Püthe, and Matt Berriman (2019), "Genedb and wikidata." *Wellcome open research*, 4, 114.

McDonnell, Erin, Kimchi Strasser, and Adrian Tsang (2018), "Manual gene curation and functional annotation." In *Fungal Genomics*, 185–208, Springer.

McEntyre, Jo and Jim Ostell (2002), "The ncbi handbook." *Bethesda (MD): National Center for Biotechnology Information (US)*.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, Lorna J Richardson, et al. (2021), "Pfam: The protein families database in 2021." *Nucleic acids research*, 49, D412–D419.

Mozannar, Hussein and David Sontag (2020), "Consistent estimators for learning to defer to an expert." In *International Conference on Machine Learning*, 7076–7087, PMLR.

Müller, H-M, Kimberly M Van Auken, Yuling Li, and Paul W Sternberg (2018), "Textpresso central: a customizable platform for searching, text mining, viewing, and curating biomedical literature." *BMC bioinformatics*, 19, 1–16.

Müller, Heiko, Felix Naumann, and Johann-Christoph Freytag (2003), "Data quality in genome databases." In *Proceedings of the Eighth International Conference on Information Quality*.

Mungall, CJ, S Misra, BP Berman, J Carlson, E Frise, N Harris, B Marshall, S Shu, JS Kaminker, SE Prochnik, et al. (2002), "An integrated computational pipeline and database to support whole-genome sequence annotation." *Genome Biology*, 3, 1–11.

Navidi, Neda and Rene Landry Jr (2021), "New approach in human-ai interaction by reinforcement-imitation learning." *Applied Sciences*, 11, 3068.

O'Leary, Nuala A, Mathew W Wright, J Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, et al. (2016), "Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation." *Nucleic acids research*, 44, D733–D745.

Olson, Nathan D, Justin Wagner, Jennifer McDaniel, Sarah H Stephens, Samuel T Westreich, Anish G Prasanna, Elaine Johanson, Emily Boja, Ezekiel J Maier, Omar Serang, et al. (2022), "Precisionfda truth challenge v2: Calling variants from short and long reads in difficult-to-map regions." *Cell Genomics*, 2, 100129.

Patterson, Robert Earl (2017), "Intuitive cognition and models of human–automation interaction." *Human factors*, 59, 101–115.

Paysan-Lafosse, Typhaine, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. (2023), "Interpro in 2022." *Nucleic Acids Research*, 51, D418–D427.

Pinski, Marc, Martin Adam, and Alexander Benlian (2023), "Ai knowledge: Improving ai delegation through human enablement." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17.

Poelchau, Monica, Christopher Childers, Gary Moore, Vijaya Tsavatapalli, Jay Evans, Chien-Yueh Lee, Han Lin, Jun-Wei Lin, and Kevin Hackett (2015), "The i5k workspace@ nal—enabling genomic data access, visualization and curation of arthropod genomes." *Nucleic acids research*, 43, D714–D719.

Ramsey, Jolene, Brenley McIntosh, Daniel Renfro, Suzanne A Aleksander, Sandra LaBonte, Curtis Ross, Adrienne E Zweifel, Nathan Liles, Shabnam Farrar, Jason J Gill, et al. (2021), "Crowdsourcing biocuration: the community assessment of community annotation with ontologies (cacao)." *PLoS computational biology*, 17, e1009463.

Ramsey, Jolene, Helena Rasche, Cory Maughmer, Anthony Criscione, Eleni Mijalis, Mei Liu, James C Hu, Ry Young, and Jason J Gill (2020), "Galaxy and apollo as a biologist-friendly interface for high-quality cooperative phage genome annotation." *PLOS Computational Biology*, 16, e1008214.

Ramstein, Guillaume P and Edward S Buckler (2022), "Prediction of evolutionary constraint by genomic annotations improves functional prioritization of genomic variants in maize." *Genome Biology*, 23, 183.

Rödelsperger, Christian, Marina Athanasouli, Maša Lenuzzi, Tobias Theska, Shuai Sun, Mohannad Dardiry, Sara Wighard, Wen Hu, Devansh Raj Sharma, and Ziduan Han (2019), "Crowdsourcing and the feasibility of manual gene annotation: a pilot study in the nematode pristionchus pacificus." *Scientific Reports*, 9, 18789.

Rutherford, Kim, Julian Parkhill, James Crook, Terry Horsnell, Peter Rice, Marie-Adèle Rajandream, and Bart Barrell (2000), "Artemis: sequence visualization and annotation." *Bioinformatics*, 16, 944–945.

Rutherford, Kim M, Midori A Harris, Antonia Lock, Stephen G Oliver, and Valerie Wood (2014), "Canto: an online tool for community literature curation." *Bioinformatics*, 30, 1791–1792.

Salamov, Asaf A and Victor V Solovyev (2000), "Ab initio gene finding in drosophila genomic dna." *Genome research*, 10, 516–522.

Salisbury, Alicia and Philippos K Tsourkas (2019), "A method for improving the accuracy and efficiency of bacteriophage genome annotation." *International journal of molecular sciences*, 20, 3391.

Sapoval, Nicolae, Amirali Aghazadeh, Michael G Nute, Dinler A Antunes, Advait Balaji, Richard Baraniuk, CJ Barberan, Ruth Dannenfelser, Chen Dun, Mohammadamin Edrisi, et al. (2022), "Current progress and open challenges for applying deep learning across the biosciences." *Nature Communications*, 13, 1728.

Sayers, Eric W, Mark Cavanaugh, Karen Clark, James Ostell, Kim D Pruitt, and Ilene Karsch-Mizrachi (2020), "Genbank." *Nucleic acids research*, 48, D84–D86.

Scalzitti, Nicolas, Anne Jeannin-Girardon, Pierre Collet, Olivier Poch, and Julie D Thompson (2020), "A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms." *BMC genomics*, 21, 1–20.

Schnoes, Alexandra M, Shoshana D Brown, Igor Dodevski, and Patricia C Babbitt (2009), "Annotation error in public databases: misannotation of molecular function in enzyme superfamilies." *PLoS computational biology*, 5, e1000605.

Schoonderwoerd, Tjeerd AJ, Emma M Van Zoelen, Karel van den Bosch, and Mark A Neerincx (2022), "Design patterns for human-ai co-learning: A wizard-of-oz evaluation in an urban-search-and-rescue task." *International Journal of Human-Computer Studies*, 164, 102831.

Shipley, Bill (2016), *Cause and correlation in biology: a user's guide to path analysis, structural equations and causal inference with R*. Cambridge university press.

Solovyev, Victor, Peter Kosarev, Igor Seledsov, and Denis Vorobyev (2006), "Automatic annotation of eukaryotic genes, pseudogenes and promoters." *Genome biology*, 7, 1–12.

Stanke, Mario, Oliver Keller, Irfan Gunduz, Alec Hayes, Stephan Waack, and Burkhard Morgenstern (2006), "Augustus: ab initio prediction of alternative transcripts." *Nucleic acids research*, 34, W435–W439.

Stanke, Mario, Stephan Waack, et al. (2003), "Gene prediction with a hidden markov model and a new intron submodel." *Bioinformatics-Oxford*, 19, 215–225.

Stein, Lincoln (2001), "Genome annotation: from sequence to biology." *Nature reviews genetics*, 2, 493–503.

Talukder, Amlan, Clayton Barham, Xiaoman Li, and Haiyan Hu (2021), "Interpretation of deep learning in genomics and epigenomics." *Briefings in Bioinformatics*, 22, bbaa177.

Tchechmedjiev, Andon, Amine Abdaoui, Vincent Emonet, Soumia Melzi, Jitendra Jonnagaddala, and Clement Jonquet (2018), "Enhanced functionalities for annotating and indexing clinical text with the ncbo annotator+." *Bioinformatics*, 34, 1962–1965.

Tello-Ruiz, Marcela K, Cristina F Marco, Fei-Man Hsu, Rajdeep S Khangura, Pengfei Qiao, Sirjan Sapkota, Michelle C Stitzer, Rachael Wasikowski, Hao Wu, Junpeng Zhan, et al. (2019), "Double triage to identify poorly annotated genes in maize: The missing link in community curation." *PLoS One*, 14, e0224086.

The UniProt Consortium (2023), "Uniprot: the universal protein knowledgebase in 2023." *Nucleic Acids Research*, 51, D523–D531.

Tiwari, Shrish, S Ramachandran, Alok Bhattacharya, Sudha Bhattacharya, and Ramakrishna Ramaswamy (1997), "Prediction of probable genes by fourier analysis of genomic sequences." *Bioinformatics*, 13, 263–270.

Vallenet, David, Laurent Labarre, Zoe Rouy, Valerie Barbe, Stephanie Bocs, Stephane Cruveiller, Aurelie Lajus, Geraldine Pascal, Claude Scarpelli, and Claudine Medigue (2006), "Mage: a microbial genome annotation system supported by synteny results." *Nucleic acids research*, 34, 53–65.

van Baren, Marijke J and Michael R Brent (2006), "Iterative gene prediction and pseudogene removal improves genome annotation." *Genome research*, 16, 678–685.

van den Bosch, Karel, Tjeerd Schoonderwoerd, Romy Blankendaal, and Mark Neerincx (2019), "Six challenges for human-ai co-learning." In *Adaptive Instructional Systems: First International Conference, AIS 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings 21*, 572–589, Springer.

van der Wal, Douwe, Iny Jhun, Israa Laklouk, Jeff Nirschl, Lara Richer, Rebecca Rojansky, Talent Theparee, Joshua Wheeler, Jörg Sander, Felix Feng, et al. (2021), "Biological data annotation via a human-augmenting ai-based labeling system." *NPJ Digital Medicine*, 4, 145.

Wang, Zifeng, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister (2022), "Learning to prompt for continual learning." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 139–149.

Watson, Joseph L, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. (2023), "De novo design of protein structure and function with rfdiffusion." *Nature*, 620, 1089–1100.

Wei, Chih-Hsuan, Alexis Allot, Robert Leaman, and Zhiyong Lu (2019), "Pubtator central: automated concept annotation for biomedical full text articles." *Nucleic acids research*, 47, W587–W593.

Wilkerson, Matthew D, Shannon D Schlueter, and Volker Brendel (2006), "yrgate: a web-based gene-structure annotation tool for the identification and dissemination of eukaryotic genes." *Genome Biology*, 7, 1–11.

Yoo, Danny, Iris Xu, Tanya Z Berardini, Seung Yon Rhee, Vijay Narayanasamy, and Simon Twigger (2006), "Pubsearch and pubfetch: a simple management system for semiautomated retrieval and annotation of biological information from the literature." *Current Protocols in Bioinformatics*, 13, 9–7.

Zhang, Qiang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. (2024), "Scientific large language models: A survey on biological & chemical domains." *arXiv preprint arXiv:2401.14656*.

Zheng, Nan-ning, Zi-yi Liu, Peng-ju Ren, Yong-qiang Ma, Shi-tao Chen, Si-yu Yu, Jian-ru Xue, Ba-dong Chen, and Fei-yue Wang (2017), "Hybrid-augmented intelligence: collaboration and cognition." *Frontiers of Information Technology & Electronic Engineering*, 18, 153–179.

Zhou, Zhongliang, Mengxuan Hu, Mariah Salcedo, Nathan Gravel, Wayland Yeung, Aarya Venkat, Dongliang Guo, Jielu Zhang, Natarajan Kannan, and Sheng Li (2023), "Xai meets biology: A comprehensive review of explainable ai in bioinformatics applications." *arXiv preprint arXiv:2312.06082*.

Zou, James, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti (2019), "A primer on deep learning in genomics." *Nature genetics*, 51, 12–18.