Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Emerging Use of Agentic AI Systems Across Genomics and Transcriptomics Domains: a Systematic Review

Irina Rozen
Bruce and Ruth Rappaport Faculty of Medicine, Technion Israel Institute of Technology, Haifa, Israel
Mahmud Omar
BRIDGE GenAI Lab, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA https://orcid.org/0009-0001-0438-0827
Yiftach Barash
BRIDGE GenAI Lab, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA https://orcid.org/0000-0002-7242-1328
Jonathan B Kruskal
Department of Radiology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
Olga R Brook
Department of Radiology, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA
Eyal Klang
BRIDGE GenAI Lab, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA https://orcid.org/0000-0002-4567-3108
Alon Gorenshtein
agorensh@bidmc.harvard.edu

BRIDGE GenAI Lab, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA https://orcid.org/0009-0000-7542-8608

Systematic Review

Keywords: Genomic, AI agent, Large language model, systematic review, AI

Posted Date: March 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9098093/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Background

Genomics and transcriptomics workflows require coordinated use of multiple specialized tools, creating technical barriers for many domain scientists. Large language models (LLMs) have shown promise for supporting bioinformatics tasks, but standalone models lack persistent state, autonomous tool use, and reliable multi-step execution. Agentic AI systems, which extend LLMs with planning, tool integration, and iterative execution, may address these limitations, yet systematic evidence of their application in genomics and transcriptomics remains limited.

Methods

We conducted a systematic review following PRISMA guidelines (PROSPERO: CRD420261292811), searching PubMed, Embase, and Web of Science. Eligible studies included LLM-based agentic systems defined by autonomous multi-step planning, iterative decision-making, and external tool or workflow invocation applied to genomics or transcriptomics tasks with formal performance evaluation. Two independent reviewers (I.R., A.G.) screened 2,932 unique records.

Results

Ten studies (2024–2026) met inclusion criteria, covering applications such as single-cell RNA-seq annotation, CRISPR guide design, Mendelian randomization, biomarker discovery, and automated bioinformatics workflows. Six systems used single-agent and four multi-agent architectures, all integrating external tools via code execution, retrieval-augmented generation, or domain-specific APIs. GPT-4–family models were the most common backbone (n = 8). Quantitative evaluations (n = 6) reported performance gains of + 1.4 to + 80 percentage points over baseline LLMs or expert comparators, while qualitative assessments (n = 4) showed high agreement with expert benchmarks. Three architectural patterns emerged: multi-agent designs were associated with tasks involving interpretive uncertainty; structural constraints reduced backbone sensitivity more effectively than model upgrades; and excessive iterative self-revision produced diminishing returns. Risk of bias was high in 80% of studies, primarily due to small datasets, lack of external validation, and subjective reference standards.

Conclusions

Agentic AI systems represent a shift from passive text generation to autonomous analytical orchestration. Be it as it may, the evidence base remains small and methodologically limited, with no system validated outside its originating research group. Future development should prioritize rigorous external benchmarking on real-world datasets, modular and explainable architectures, and coverage of underrepresented domains including variant interpretation and spatial transcriptomics.

1. Introduction

Genomics and transcriptomics are central to modern biological research, yet their analytical pipelines are technically demanding, requiring coordinated execution of specialized tools across multiple sequential steps.[1,2] A typical workflow spans quality control, read alignment, quantification, statistical testing, and functional enrichment analysis, each stage with its own software requirements, parameters, and conventions.[3–6] Navigating this process is time-consuming and requires substantial training, and even experienced bioinformaticians face repetitive overhead when adapting pipelines to new datasets or experimental designs.[7,8]

The availability of large language models (LLMs) has prompted interest in their potential to assist with bioinformatics tasks through text-based interaction, including methodological guidance, interpretation of scientific literature, and support for biological reasoning.[9–12] However, base LLMs were not built for multi-step procedural work: they lack persistent state across steps, cannot invoke external tools directly, and often produce unreliable outputs when tasks require combining domain knowledge with precise computational execution.[13,14]

Agentic AI systems address some of these limitations by extending LLM capabilities through autonomous planning, iterative execution, and interaction with external tools and data sources, thereby supporting greater automation of analytical workflows.[15–19] Multi-agent

designs distribute reasoning across specialized agents that communicate and validate outputs collaboratively, which is particularly relevant where task complexity or interpretive judgment exceeds what a single model can reliably handle.[20–24] Despite growing interest, systematic evidence on the application of agentic AI in genomics and transcriptomics remains limited.

This systematic review evaluates the current state of agentic AI across these domains, characterizes the architectural and technical approaches employed, and identifies the biological tasks and design considerations most relevant to this emerging field.

2. Methods

Protocol and registration

We conducted this review using an established methodological framework consistent with the extended PRISMA statement.[25] The study protocol is available on PROSPERO (CRD420261292811).

Data Sources and Searches

We performed a systematic literature search across three primary bibliographic databases—PubMed, Embase, and Web of Science without time limitation. The search was designed to identify studies at the intersection of agentic AI and genomic/transcriptomics domains in the field of medicine and biology, utilizing a combination of Medical Subject Headings (MeSH), Emtree terms, and free-text keywords.

The search strategy was structured around three core conceptual domains: (1) large language models and generative AI, (2) agentic and autonomous AI frameworks, and (3) genomics and transcriptomics. Detailed, reproducible search strings for each database are provided in the Supplementary Materials. Forward and backward citation tracking was performed to ensure comprehensive coverage.

Study screening

To ensure the review's methodological integrity, we applied predefined inclusion and exclusion criteria during the screening process to isolate high-quality literature implementing agentic AI systems in genomics and transcriptomics.

We specifically sought out studies describing agentic AI architectures, defined as an LLM-based system that autonomously plan, execute, and adapt multi-step tasks, including iterative decision-making and external tool or workflow invocation, to achieve a specified objective. Furthermore, we restricted the scope to the domains of genomics and transcriptomics, targeting tasks such as, but not limited to, sequence analysis, variant interpretation, gene expression profiling, and single-cell RNA sequencing analysis. Each included study was required to provide formal validation, whether through quantitative performance metrics or systematic qualitative expert evaluation. Only peer-reviewed journal articles published in English were considered.

Conversely, we excluded research focused on non-agentic AI, such as standard deep learning models or "single-pass" LLM applications that lack iterative loops and autonomous planning, consistent with prior definitions of agentic AI.[15] Purely descriptive or visionary papers, including editorials and perspectives that did not present an original system implementation, were also excluded. While the review is bioinformatic in nature, we excluded studies focused exclusively on proteomics, metabolomics, or protein folding unless they were embedded within a broader genomic or transcriptomic workflow. Finally, to maintain transparency and reproducibility, we excluded conference abstracts, preprints and papers where the underlying architecture or methodology was not fully disclosed.

Two reviewers (I.R, A.G) independently screened all titles and abstracts using Rayyan systematic review software. Articles marked for potential inclusion by either reviewer underwent full-text assessment. Disagreements were resolved through discussion, with a third reviewer (E.K) providing arbitration when consensus could not be reached.

Data extraction

We developed a standardized data extraction form and refined it during a pilot phase that included five studies. Two reviewers (I.R, A.G) independently extracted data, with disagreements resolved through discussion and consensus. Data were extracted using a standardized set of pre-specified elements to allow consistent comparison across studies. We collected basic study characteristics, including publication year, system name, application domain, and the biological task(s) addressed.

For each system, we documented technical features such as the base LLM, agentic architecture, implementation framework, and how external tools were integrated into the workflow. To evaluate performance, we extracted the primary outcome reported by the authors, the comparator used, and the relative performance of the agent compared with that comparator. We also documented the evaluation sample size and datasets, as well as whether synthetic data were used and whether independent external validation was performed.

Details regarding abstraction rules and counting conventions are provided in the Supplementary Materials.

Risk of bias assessment

We adapted the QUADAS-AI protocol[26] to assess the risk of bias and applicability concerns for the evaluated AI agents. This modified instrument evaluated four domains: (1) Patient/Data Selection, (2) Index Test (the AI agent), (3) Reference Standard, and (4) Flow and Timing. Each domain was rated as low risk, high risk, or unclear risk of bias. Additionally, we assessed concerns regarding applicability for the first three domains.

Specific high-risk indicators included: reliance on synthetic or simulated data distributions rather than real-world biological datasets; use of small, curated convenience samples ($n < 100$) likely to introduce selection bias; lack of independent external validation cohorts; reliance on subjective expert consensus or LLM-generated outputs as the reference standard instead of experimental or database-derived ground truth; and selective outcome reporting where agent failures were excluded from final metrics. Two reviewers (I.R., A.G.) independently assessed each study, with disagreements resolved through third reviewer (E.K.).

Synthesis of findings

Given the variability in system design, biological focus, evaluation methods, and reporting practices, a quantitative meta-analysis was not appropriate. We therefore conducted a qualitative analysis guided by the predefined research questions of this review.

Studies were first organized according to agentic architecture, distinguishing between single-agent and multi-agent systems based on how reasoning and control were structured. We then examined patterns of tool use, including code or workflow execution, structured database or API querying, web search and literature retrieval, and retrieval-augmented generation (RAG).

To better understand how these design choices relate to application context, we further grouped studies by biological domain and task. These included genomics, transcriptomics—particularly single-cell RNA sequencing—and multi-omics integration, as well as specific objectives such as statistical association testing, cell-type annotation, gene-editing design, and functional interpretation.

3. Results

Study selection

Our systematic search across PubMed, Embase, and Web of Science identified an initial 3,393 records. Following the removal of 461 duplicates, 2,932 unique citations remained for title and abstract screening. From this pool, 116 reports were sought for retrieval and underwent a full-text eligibility assessment. Ultimately, 18 studies satisfied all inclusion criteria and were incorporated into the final review (Fig. 1). During the full-text evaluation, reports were excluded if they focused on non-genomic or non-transcriptomic domains ($n = 70$), were review articles rather than primary research ($n = 19$), or lacked comparative performance metrics ($n = 9$).

Overview of the included studies

A total of 18 articles met the predefined inclusion criteria. However, only 10 studies were included in the final discussion, as nine articles described extensions of the same underlying framework (AI-HOPE), reporting domain-specific refinements, or expanded use cases. To avoid overweighting a single system, only the first article reporting the initial system description and benchmark evaluation was included, while subsequent related publications are summarized in the Appendix.

The 10 included studies, published between 2024 and 2026, focused on genomics, transcriptomics, and multi-omics applications, encompassing a range of biological tasks including scRNA-seq cell type annotation, biomarker discovery, CRISPR guide design, Mendelian randomization, and end-to-end bioinformatics workflow automation (Table 1).

Architecturally, single-agent systems were most common (n = 6), while four studies utilized multi-agent frameworks to manage supervisory or task-specialized roles. All the systems utilized external tools, with their integration mechanisms most commonly relying on code executors, RAG, or domain-specific web APIs (Table 1). The GPT family was the most common backbone for these agents (n = 8), although specialized or smaller models like LLaMA-3 and Phi-3 were also successfully deployed (**Supplementary Table 1**).

Evaluation datasets varied in both scale and unit of analysis, including gene sets (n ≈ 886–1,106), cell populations (n = 970), RNA-seq samples (n = 8), exposure–outcome pairs (n = 40), sets of academic papers (n = 272 articles), benchmark CRISPR design test cases (n = 288), real-world bioinformatics analysis cases (n = 40), and large public scRNA-seq collections (~ 1 million cells across 25 studies) (Table 1).

Across the included studies, agentic systems were evaluated against base LLMs (n = 6) or human domain experts (n = 3), or established literature benchmarks (n = 1), using task-specific performance measures appropriate to each application. Reported outcomes clustered into quantitative performance improvements over baseline comparators (n = 6), expressed as absolute gains in task accuracy, success rate, or similarity, and qualitative performance, agreement, or equivalence assessments (n = 4), reflecting validation against expert judgment, workflow concordance, or established analytical results. Among studies reporting quantitative comparisons, absolute performance improvements ranged from + 1.4 to + 80 percentage points, whereas qualitative evaluations reported performance equivalence or high agreement with expert or literature benchmarks (**Supplementary Table 1**).

Table 1
Study Characteristics of Included Agentic AI Systems in Genomics and Transcriptomics

Study (Year)	System	Domain	Biological Task(s)	Agentic Architecture	Framework	Tool Integration	Evaluation Sample
Yang & Velazquez-Villarreal (2025)	AI-HOPE	Oncology / Genomics	Clinical–genomic association and survival analysis	Multi-agent (1 supervisor agent + 2 analytical agents: group analysis, data exploration)	Custom	n = 1: Code / workflow execution (statistical analysis and hypothesis testing)	2 analytical studies
Qu et al. (2025)	CRISPR-GPT	Genomics	CRISPR guide design and experimental analysis	Multi-agent (n = 4)	Custom	n = 4: (1) Web search & literature retrieval; (2) Structured database/API querying (Primer3, CRISPRitz, CRISPResso2); (3) Code / workflow execution; (4) RAG	288 benchmark test cases
Xie et al. (2025)	CASSIA	Transcriptomics	scRNA-seq cell type annotation	Multi-agent (5 core + up to 4 optional)	Custom (Python/R)	n = 1: RAG (cell-marker references)	970 cell populations
Nouri et al. (2026)	CellAtria	Transcriptomics	End-to-end scRNA-seq data ingestion and analysis	Single agent	LangGraph	n = 1: Code / workflow execution (CellExpress pipeline for scRNA-seq QC, normalization, clustering)	25 public scRNA-seq studies (~ 1M post-QC cells)
Arowolo et al. (2025)	ARMOA	Oncology/Multi-omics (genomics + transcriptomics + proteomics + metabolomics)	(1) Biomarker discovery related to PI3K/AKT dysregulation; (2) Drug repurposing candidate prioritization	Single agent with modular analytical components (GNN + ML)	Custom (Python)	n = 1: RAG (literature sources, curated biological and pharmacological knowledge bases)	1,000 samples, each with 400 features
Zhou et al. (2024)	AutoBA	Multi-omics (genomics, transcriptomics, proteomics, metabolomics)	End-to-end conventional omics analyses (e.g., RNA-seq differential expression, scRNA-seq clustering, ChIP-seq peak calling, WGS/WES variant calling)	Single agent	Custom (Python)	n = 1: Code / workflow execution (automated bioinformatics pipeline orchestration)	40 real-world analysis cases
Mehandru et al. (2025)	BioAgents	Genomics + Transcriptomics	Bioinformatics workflow design and code/workflow generation	Multi-agent system (n = 3)	Custom (Python)	n = 2: (1) RAG (nf-core modules, workflow documentation); (2) Tool metadata/ontology retrieval (Sequence Ontology, software metadata)	3 workflows × 2 task types = 6 evaluation tasks
Pickard et al. (2025)	BRAD	Genomics + Transcriptomics	Biomarker discovery and	Single agent	LangChain	n = 3: (1) RAG (scientific	886 gene sets

Study (Year)	System	Domain	Biological Task(s)	Agentic Architecture	Framework	Tool Integration	Evaluation Sample
			enrichment analysis			literature); (2) Structured database/API querying (Gene Ontology, Enrichr); (3) Code/workflow execution (external analysis software)	(enrichment); 8 RNA-seq samples (case study)
Wang et al. (2025)	GeneAgent	Genomics + Transcriptomics	Identification of the main biological process shared by a given list of genes	Single agent	Custom (Python)	n = 2: (1) Structured database/API querying (g:Profiler, Enrichr, NCBI E-utils); (2) Web search & literature retrieval	1,106 gene sets + 7 novel mouse gene sets
Xu et al. (2025)	MRAgent	Genomics	Automated discovery and causal validation of exposure–outcome relationships using Mendelian Randomization	Single agent	Custom (Python)	n = 2: (1) Web search & literature retrieval (biomedical publications); (2) Structured database/API querying (curated medical evidence sources)	272 PubMed articles; 40 exposure–outcome pairs; 30 diseases; 8 MR analyses

Abbreviations: RAG, retrieval-augmented generation; GNN, graph neural network; ML, machine learning; QC, quality control; scRNA-seq, single-cell RNA sequencing; MR, Mendelian randomization; WGS, whole genome sequencing; WES, whole exome sequencing.

Tool count reflects distinct functional capabilities. Multiple databases accessed via the same interface are counted as a single tool type. Evaluation sample size refers to the unit used by the authors to evaluate system performance.

Risk of bias

Quality assessment using QUADAS-AI criteria revealed significant methodological concerns regarding internal validity and applicability. Only 20% (2/10) of studies (Xie et al., Wang et al.)[27,28] achieved low risk of bias across all domains. The remaining 80% demonstrated high risk in at least one domain, with patient selection representing the most problematic area (80% high risk). This bias largely stemmed from the use of small, curated convenience samples or specific "demonstration" cases rather than consecutive or random sampling from established clinical cohorts. Other sources of bias stemmed from: synthetic data dependence (n=1, 10%), single-center designs without external clinical validation (n=10, 100%), and reference standard concerns (n=7, 70%) due to reliance on subjective expert consensus or literature alignment rather than objective ground truth (**Supplementary Table 2**).

1. Architecture

1.1 Single-agent systems

Six studies employed a single centralized LLM to manage reasoning and tool use (**Table 1**). AutoBA (Zhou et al., 2024)[29] and CellAtria (Nouri et al., 2026)[30] used this design for end-to-end automation of bioinformatics pipelines, with CellAtria further constraining execution through a graph-based workflow for sequential data ingestion. ARMOA (Arowolo et al., 2025)[31] and BRAD (Pickard et al., 2025)[32] combined single-agent control with retrieval-based knowledge access to support biomarker discovery, while GeneAgent (Wang et al., 2025)[28] and MRAgent (Xu et al., 2025)[33] targeted narrower analytical tasks such as functional interpretation and causal inference.

Across these systems, reliability was primarily associated with constrained execution rather than open-ended reasoning. When execution was constrained through fixed tool sequences, mandatory database verification, or pre-vetted workflows, performance was stable. For example, CellAtria achieved consistent run-to-run behavior (Gini ≈ 0.10) across ~1 million cells and across LLM backends (GPT-4o versus GPT-4o-mini)[30], GeneAgent reduced hallucinated enrichment terms from 44% to 19.3% through database verification[28], and AutoBA employed automated error-recovery loops to handle execution failures without human intervention.[29] In contrast, where such constraints were absent, performance degraded: MRAgent showed wide variability across LLM backends (F1 range: 0.16–0.77).[33]

1.2 Multi-agent systems

Four studies adopted multi-agent architectures to decompose complex tasks into specialized roles. CRISPR-GPT (Qu et al., 2025) assigned distinct agents to experimental planning and guide RNA design, reflecting the multi-stage structure of gene-editing workflows.[34] CASSIA (Xie et al., 2025) employed multiple agents for iterative cell-type annotation and validation, using consensus mechanisms to reduce interpretative error.[27] BioAgents (Mehandru et al., 2025) separated biological reasoning from software retrieval[35], while AI-HOPE (Yang & Velazquez-Villarreal, 2025) used a supervisory agent to coordinate clinical–genomic exploration.[36]

The observed pattern across these systems was separation of generation and evaluation as an error-control mechanism. CASSIA's independent Scorer agent improved fully correct annotations by 12–41% over single-prompt baselines[27], and CRISPR-GPT's domain-specific sub-agent outperformed a general LLM by 12% in task-specific Q&A accuracy.[37] However, increased deliberation was not universally beneficial: both CASSIA and BioAgents explicitly capped revision cycles due to diminishing or negative returns with excessive iteration, indicating that multi-agent deliberation has an optimal, bounded depth.[35,38] (**Supplementary Figure 1**)

1.3 Tool integration

External tool integration was observed across all reviewed systems, but the type and role of tool use varied markedly. A first group of agents relied exclusively on code or workflow execution as their primary mechanism for task completion. AI-HOPE (Yang and Velazquez-Villarreal) enables execution of predefined statistical workflows to support hypothesis-driven analyses such as survival and association testing.[39] AutoBA (Zhou et al.) similarly focuses on automated execution and orchestration of bioinformatics pipelines, emphasizing end-to-end workflow management rather than textual reasoning.[29] CellAtria (Nouri et al.) was explicitly designed to operate without retrieval-based components, instead coordinating the CellExpress pipeline for standardized scRNA-seq processing through a strictly tool-mediated execution framework.[30] A second group of systems uses additional tools to ground their outputs in external biological knowledge. Wang et al. introduced GeneAgent, which queries specialized web APIs, including g:Profiler and Enrichr, to verify its gene-set interpretations against curated databases.[28] Other frameworks utilize RAG to provide the model with task-relevant contextual information. Mehandru et al.'s BioAgents applies RAG to nf-core module documentation, helping the agent select and configure the correct tools for complex bioinformatics workflows.[35]

1.4 Base LLMs

GPT-4–based models accessed via commercial APIs were the most frequently used base LLMs across the included systems (**Supplementary Table 1**), although agentic architectures were also implemented using a range of proprietary and open-weight model backbones. For example, Xu et al. (2025) evaluated MRAgent across a broad range of backbones including Claude-3, Qwen, LLaMA-3 and Mixtral[33], while Zhou et al. (2024) tested AutoBA using both online (ChatGPT-3.5, ChatGPT-4) and offline models like CodeLlama and LLaMA-2.[29]

Some studies opted to rely on locally deployable, open-weight LLMs for practical deployment reasons. Yang and Velazquez-Villarreal (2025) utilized a locally deployed LLaMA-3 model for AI-HOPE to support a closed-system setup that minimizes external data exchange, motivated by privacy and regulatory considerations in clinical-genomic settings.[39] In contrast, Mehandru et al. (2025) built BioAgents on the smaller Phi-3-mini model to reduce infrastructure demands and computational costs, enabling more accessible local deployment and real-time bioinformatics applications.[35]

Backbone sensitivity decreased when correctness was enforced by structural constraints rather than left entirely to a LLM's internal reasoning. In MRAgent, selecting relevant GWAS IDs depended heavily on model-internal semantic reasoning, resulting in substantial performance variation across LLMs (F1 range: 0.16–0.77).[33] In contrast, CellAtria explicitly hardens execution via tool-schema validation and restricted invocation patterns and then shows near-backbone invariance, maintaining similar runtime (1.45 ± 1.25 vs 1.70 ± 1.70 min) and low-dispersion outputs (Gini ≈ 0.10–0.14) when switching from GPT-4o to GPT-4o-mini.[30] Similarly, CASSIA reduced variability (accuracy 0.92–0.82) through a generator–scorer multi-agent design that separates generation from evaluation, allowing errors to be detected and refined (**Figure 2**).[38] Together, these findings suggest that backbone variability depends less on the number of agents or tools, and more on whether the system includes structured mechanisms, either validation agents or constrained tool layers, that limit reliance on unconstrained model reasoning.

2. Purpose of the model

All included systems targeted tasks within genomics, transcriptomics, or their integration with additional omics layers. In genomics and mixed domains (n=6), three multi-agent systems (n=3), CRISPR-GPT, AI-HOPE, and BioAgents, implemented explicit task decomposition or specialized agents to coordinate experimental planning, clinical–genomic association testing, or workflow/code generation accordingly.[35,37,39] Single-agent systems (n=3), GeneAgent, BRAD, and MRAgent, targeted bounded analytical aims (gene-set functional labeling, enrichment-driven biomarker interpretation, and MR-based causal inference) and increased reliability via tool- and database-grounded execution, rather than inter-agent deliberation.[28,32,33]

Transcriptomic systems centered on single-cell RNA sequencing but diverged markedly in purpose. CASSIA (Xie et al., 2025) addressed cell-type annotation, a biologically ambiguous task lacking universal ground truth, using a multi-agent framework to iteratively validate and calibrate predictions.[38] In contrast, CellAtria (Nouri et al., 2026) automated metadata extraction and dataset ingestion for reproducible pipeline execution.[30] Because its task was logistical and deterministic, CellAtria functioned effectively as a single-agent system. Thus, within the same omics layer, architecture reflected interpretive uncertainty rather than data type.

Two studies addressed multi-omics integration. ARMOA (Arowolo et al., 2025) performed pathway-centric analysis across genomic, transcriptomic, proteomic, and metabolomic layers to support biomarker discovery and drug repurposing[31], while AutoBA (Zhou et al., 2024) orchestrated general bioinformatics workflows across modalities.[29] Notably, both employed single-agent designs. This reflects that agentic architecture was driven by task-specific requirements rather than by the breadth of omics modalities alone (Figure 3).

AI-HOPE case

Following the initial AI-HOPE publication (Yang & Velazquez-Villarreal, 2025)[39], multiple domain-specialized derivatives ("AI-HOPE-X"), including pathway-focused agents (e.g., PI3K, TGFβ, WNT, JAK-STAT, RTK-RAS, MAPK, TP53)[40–46] and a population-level extension (AI-HOPE-PM)[47], were subsequently reported (Supplementary Table 3, Supplementary Figure 2). These variants retain the same conversational natural-language–to–analysis architecture and statistical backbone, differing primarily in biological scope, dataset configuration, and incremental usability refinements. A structured comparative analysis of these extensions, with detailed characteristics summarized in Supplementary Materials.

4. Discussion

This systematic review synthesizes 10 peer-reviewed agentic AI systems applied to genomics and transcriptomics. We found that AI agent architectures tend to be adapted according to the complexity of the task within each domain. More complex multi-agent frameworks are generally used for tasks where the output involves greater uncertainty, whereas simpler, deterministic tasks are typically addressed using simpler architectures such as single-agent systems. This suggests that the choice of AI agent framework is driven by task complexity, consistent with observations in prior reviews of AI agents in clinical medicine.[29]

1. Implementation Guidance

1.1 Architecture Is Driven by Task Uncertainty, Not Omics Complexity

Across the reviewed systems, architectural complexity appeared more closely associated with task uncertainty than with the number of omics modalities involved. Multi-agent designs were predominantly employed in tasks characterized by biological ambiguity and the absence of stable ground truth, such as cell type annotation[27], where iterative hypothesis refinement can reduce classification uncertainty. In contrast, tasks with deterministic analytical structures, including pipeline orchestration[30] or formal causal inference[31], were successfully implemented within single-agent frameworks. Even in multi-agent systems, deliberative depth was typically capped (both CASSIA and BioAgents)[27,32] explicitly limited revision cycles due to diminishing returns, consistent with broader evidence that repeated self-revision without external feedback yields negligible or negative gains in LLM performance.[33,34]

1.2 Structural Constraints Reduce LLM Backbone Sensitivity

Across systems evaluated on multiple LLM backbones, differences in performance were more strongly shaped by architectural design than by model choice alone. Backbone effects were most evident when correctness relied primarily on unconstrained internal reasoning, whereas systems incorporating structured validation mechanisms showed comparatively stable performance across different base LLMs. When execution was guided by schema-restricted tool invocation or by explicit separation between generation and evaluation, substituting backbones produced relatively modest variability. This pattern suggests a practical implication: in agentic bioinformatics systems, strengthening architectural safeguards may improve reproducibility more reliably than upgrading the underlying model.

This principle extends beyond backbone robustness to the broader problem of hallucination. Hallucination is a known limitation of LLMs[35] and extends beyond basic models to agentic LLM systems.[36,37] In the genomic and transcriptomic domain, we found that agentic frameworks implement a range of safeguards designed to mitigate or avoid hallucinations.[27,28,30,38,39]

1.3 Explainability and Trust

One of the problems of AI is the black box question, we are lacking the knowledge on why the model works in a certain way. One of the ways to gain the trust of researchers and physicians is to provide these models with tools that explain the way models act in a certain way to improve the trust of the users. Among reviewed systems, only ARMOA employs formal explainable AI (XAI) methods, specifically SHAP values and GNN attention weights, to attribute predictions to specific input features.[40] Several others adopt forms of procedural transparency: reasoning traces, evidence grounding, or structured natural language rationale.[27,28,32,39] These are not formal XAI in the technical sense, but they may serve a similar communicative function by making system behaviour more inspectable.[41,42] Some reviewed systems provide no dedicated interpretability mechanism at all.[41,42]

Whether any of these approaches builds trust among domain scientists remains untested in the reviewed work. None of the systems evaluated user trust empirically. Trust has been identified as a significant predictor of AI acceptance and continued use.[43]

1.4 Modular Extensibility

Several of the reviewed systems explicitly adopt modular design as an architectural principle, though this is not universal across the field. Where it is present, the motivation appears to be longevity: modular architectures allow components, databases, tools, or LLM backends, to be updated or extended without redesigning the core system. The AI-HOPE family offers one concrete illustration of this, where a shared architecture was extended into multiple domain-specific derivatives targeting distinct biological pathways.[44–52] This suggests that when modularity is deliberately built in, it can meaningfully reduce the cost of expanding scope. At the same time, modularity introduces its own challenges: as systems grow through extension, maintaining consistency in behaviour, validation standards, and documentation across components becomes non-trivial. Whether the reviewed systems have adequate mechanisms for this kind of governance is not well addressed in the primary literature and remains an open concern.

3. Decision Framework for Agentic AI Design

Synthesizing these architectural patterns, Table 2 (decision framework) maps seven binary design questions, covering target audience, task ambiguity, backbone sensitivity, data privacy, knowledge grounding, explainability and extensibility, to architectural implications grounded in the reviewed evidence.

Table 2: Decision Framework for Agentic AI Design in Genomics & Transcriptomics

Design Question	YES → Implication	NO → Implication	Evidence from Reviewed Systems
Is the intended user a domain expert without programming skills (bench scientist, clinician, student)?	Prioritise a natural-language conversational interface; minimise scripting requirements; validate with non-expert users in real workflows.	A command-line or API interface may be acceptable; reproducibility and performance benchmarks take priority over interface simplicity.	*AI-HOPE (oncologists), CRISPR-GPT (junior researchers), CellAtria & CASSIA (bench scientists), MRAgent (clinicians), BioAgents (students) all target non-programmers and provide NL interfaces.*
Does the task involve biological ambiguity or absence of stable ground truth (e.g., cell-type annotation, hypothesis generation)?	Adopt a multi-agent architecture with iterative refinement cycles; include a dedicated validation or scoring agent.	A single-agent architecture is likely sufficient; multi-agent overhead is not justified for deterministic pipelines.	*CASSIA and BioAgents use multi-agent designs for annotation/workflow guidance; CellAtria and MRAgent use single-agent designs for deterministic pipeline execution and causal inference.*
Is consistent reproducibility across LLM backbone versions a priority (e.g., clinical or regulated setting)?	Implement structural constraints: schema-restricted tool invocation, separation of generation and evaluation, pre-validated pipelines. Architecture should absorb backbone variability.	If reproducibility is less critical and flexibility is valued, lighter validation mechanisms may be acceptable, but hallucination risk should still be addressed.	*CellAtria's schema-restricted tools showed stable behaviour across LLM versions. CASSIA's independent scoring agent reduce backbone sensitivity.*
Does the system handle patient-level, clinical, or otherwise regulated genomic data?	Evaluate HIPAA-compliant cloud endpoints, local/on-premise deployment with open-weight models, sequence-level input filtering, or federated computation with differential privacy.	Cloud-based deployment with standard API access is likely permissible; standard data handling practices apply.	*GeneAgent uses HIPAA-compliant endpoints; AI-HOPE and AutoBA support local open-weight models; CRISPR-GPT applies sequence-level filtering; ARMOA implements federated + differential privacy.*
Does the task depend on domain knowledge (databases, literature, guidelines)?	Implement RAG over curated, regularly updated domain corpora; monitor retrieval corpus quality; establish update schedules.	Parametric LLM knowledge may be sufficient for stable domains	*CRISPR-GPT, CASSIA, and ARMOA use RAG over domain databases; BioAgents used RAG over nf-core documentation, but errors were attributed to gaps in the indexed corpus, illustrating retrieval quality as a limiting factor.*
Does the system operate in a high-stakes context where users must understand and justify AI-generated outputs (clinical management, experimental design)?	Implement explainability mechanisms: reasoning traces, evidence grounding, quality scores, or — where predictions must be attributed to input features — formal XAI (e.g., SHAP, attention weights).	Explainability infrastructure may be less critical for exploratory or low-stakes tasks; standard output logging may suffice.	*Only ARMOA employs formal XAI (SHAP + GNN attention). CASSIA, BRAD, BioAgents, and GeneAgent provide procedural transparency. No reviewed system empirically validated whether these mechanisms improved user trust.*
Is the system intended to support multiple biological domains, tasks, or evolving LLM backends over time?	Design for modularity from the outset: encapsulate tools as atomic functions, decouple the LLM orchestration layer from domain logic, document component I/O contracts.	A single-purpose, tightly integrated design may be more efficient if the scope is fixed and long-term extension is not anticipated.	*The AI-HOPE architecture was extended into 8+ pathway-specific derivatives. CRISPR-GPT's three-step integration protocol (tool wrapping, meta-mode, auto-mode registration) allows addition of new capabilities without core redesign.*

5. Limitations

The evidence base is small (n = 10) and methodologically heterogeneous, precluding quantitative meta-analysis. Quality assessment using QUADAS-AI criteria revealed high risk of bias in 80% of systems, most commonly due to small, curated datasets and absence of independent external validation; none of the systems were evaluated outside their originating research group. Benchmarking against base LLMs or in-group human experts may overstate the added value of the agentic architecture. The rapid pace of LLM development means comparisons to specific model versions may understate current baselines. Important genomics and transcriptomics tasks, like variant interpretation, epigenomics, spatial transcriptomics, are absent from the evidence base, limiting generalizability.

6. Conclusions

Agentic AI transforms genomics and transcriptomics by transitioning from simple text generation to autonomous, multi-step analytical orchestration. By integrating LLMs with external tools, databases, and multi-agent reasoning, these systems lower programming barriers for domain specialists and enhance the reproducibility of complex workflows. Our findings indicate that architectural complexity is primarily driven by task uncertainty rather than data breadth, with structural constraints effectively mitigating backbone sensitivity and

hallucinations. However, significant challenges remain, including high risks of methodological bias, a lack of independent external validation, and limited coverage of emerging fields like spatial transcriptomics. Future research must prioritize rigorous benchmarking in real-world experimental settings and develop modular, explainable frameworks to ensure these agents evolve from prototypes into trusted, transparent collaborators in the field of biomedical research.

Declarations

Author Contributions: Conceptualization, IR, AG, EK, MO, GN; Methodology, IR, AG, EK; Formal Analysis, IR, AG; Data Curation, IR, AG; Writing-Original Draft Preparation, AG,IR; Writing- Review & Editing, IR, AG, MO, EK,YB, JBK, ORB; Supervision: EK

Funding. None

Competing interests. The authors declare that they have no competing interests.

References

1. Stark R, Grzelak M, Hadfield J. RNA sequencing: the teenage years. Nat Rev Genet. 2019;20(11):631-656. doi:10.1038/s41576-019-0150-2
2. Gauthier J, Vincent AT, Charette SJ, Derome N. A brief history of bioinformatics. Accessed February 17, 2026. https://dx.doi.org/10.1093/bib/bby063
3. Conesa A, Madrigal P, Tarazona S, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(1):13. doi:10.1186/s13059-016-0881-8
4. Dawadi P, Pokharel B, Shrestha A, et al. From bench to bytes: a practical guide to RNA sequencing data analysis. Front Genet. 2025;16:1697922. doi:10.3389/fgene.2025.1697922
5. Auwera GAV der, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics / editoral board, Andreas D Baxevanis . [et al]. 2013;11(1110):11.10.1. doi:10.1002/0471250953.bi1110s43
6. Fc G, H M, L K, Mr C. Chromatin accessibility profiling by ATAC-seq. Nature protocols. 2022;17(6). doi:10.1038/s41596-022-00692-9
7. Ewels PA, Peltzer A, Fillinger S, et al. The nf-core framework for community-curated bioinformatics pipelines. Nat Biotechnol. 2020;38(3):276-278. doi:10.1038/s41587-020-0439-x
8. L W, A W, J G. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nature methods. 2021;18(10). doi:10.1038/s41592-021-01254-9
9. Lin A, Ye J, Qi C, et al. Bridging artificial intelligence and biological sciences: a comprehensive review of large language models in bioinformatics. Briefings in Bioinformatics. 2025;26(4):bbaf357. doi:10.1093/bib/bbaf357
10. Liu J, Yang M, Yu Y, et al. Advancing bioinformatics with large language models: components, applications and perspectives. arXiv. Preprint posted online January 31, 2025:arXiv:2401.04155. doi:10.48550/arXiv.2401.04155
11. Alam K, Roy B. From Prompt to Pipeline: Large Language Models for Scientific Workflow Development in Bioinformatics. arXiv. Preprint posted online August 14, 2025:arXiv:2507.20122. doi:10.48550/arXiv.2507.20122
12. Ruan W, Lyu Y, Zhang J, et al. Large language models for bioinformatics. Quant Biol. 2026;14(1):e70014. doi:10.1002/qub2.70014
13. Shan L, Luo S, Zhu Z, Yuan Y, Wu Y. Cognitive Memory in Large Language Models. arXiv. Preprint posted online April 24, 2025:arXiv:2504.02441. doi:10.48550/arXiv.2504.02441
14. Yang T, Xiao Y, Bao Z, Hao J, Peng J. The rise and potential opportunities of large language model agents in bioinformatics and biomedicine. Brief Bioinform. 2025;26(6). doi:10.1093/bib/bbaf601
15. Wang L, Ma C, Feng X, et al. A survey on large language model based autonomous agents. Front Comput Sci. 2024;18(6):186345. doi:10.1007/s11704-024-40231-1
16. Krishnan N. AI Agents: Evolution, Architecture, and Real-World Applications. arXiv. Preprint posted online March 16, 2025:arXiv:2503.12687. doi:10.48550/arXiv.2503.12687
17. Yang Y, Chai H, Song Y, et al. A Survey of AI Agent Protocols. arXiv. Preprint posted online June 21, 2025:arXiv:2504.16736. doi:10.48550/arXiv.2504.16736
18. Li X. A Review of Prominent Paradigms for LLM-Based Agents: Tool Use (Including RAG), Planning, and Feedback Learning. arXiv. Preprint posted online 2024. doi:10.48550/ARXIV.2406.05804

19. Zhou J, Jiang J, Han Z, Wang Z, Gao X. Streamline automated biomedical discoveries with agentic bioinformatics. Brief Bioinform. 2025;26(5). doi:10.1093/bib/bbaf505
20. Du Y, Li S, Torralba A, Tenenbaum JB, Mordatch I. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv. Preprint posted online 2023. doi:10.48550/ARXIV.2305.14325
21. Guo T, Chen X, Wang Y, et al. Large Language Model based Multi-Agents: A Survey of Progress and Challenges. arXiv. Preprint posted online 2024. doi:10.48550/ARXIV.2402.01680
22. Zou J, Topol EJ. The rise of agentic AI teammates in medicine. The Lancet. 2025;405(10477):457. doi:10.1016/S0140-6736(25)00202-8
23. Tsiotras P, Gombolay M, Foerster J. Editorial: Decision-making and planning for multi-agent systems. Front Robot AI. 2024;11:1422344. doi:10.3389/frobt.2024.1422344
24. Sapkota R, Roumeliotis KI, Karkee M. AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges. arXiv.org. May 15, 2025. doi:10.1016/j.inffus.2025.103599
25. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. Published online March 29, 2021. doi:10.1136/bmj.n71
26. Guni A, Sounderajah V, Whiting P, Bossuyt P, Darzi A, Ashrafian H. Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies Using AI (QUADAS-AI): Protocol for a Qualitative Study. JMIR Research Protocols. 2024;13:e58202. doi:10.2196/58202
27. Xie E, Cheng L, Shireman J, et al. CASSIA: a multi-agent large language model for automated and interpretable cell annotation. Nat Commun. 2025;17(1):389. doi:10.1038/s41467-025-67084-x
28. Wang Z, Jin Q, Wei CH, et al. GeneAgent: Self-verification Language Agent for Gene Set Knowledge Discovery using Domain Databases. Preprint posted online May 25, 2024.
29. Zhou J, Zhang B, Li G, et al. An AI Agent for Fully Automated Multi-Omic Analyses. Advanced Science. 2024;11(44):2407094. doi:10.1002/advs.202407094
30. Nouri N, Artzi R, Savova V. An agentic AI framework for ingestion and standardization of single-cell RNA-seq data analysis. npj Artif Intell. 2026;2(1):8. doi:10.1038/s44387-025-00064-0
31. Arowolo MO, Abdulsalam SO, Isiaka RM, et al. Agentic RAG-Driven Multi-Omics Analysis for PI3K/AKT Pathway Deregulation in Precision Medicine. Algorithms. 2025;18(9):545. doi:10.3390/a18090545
32. Pickard J, Prakash R, Choi MA, et al. Automatic biomarker discovery and enrichment with BRAD. Wren J, ed. Bioinformatics. 2025;41(5):btaf159. doi:10.1093/bioinformatics/btaf159
33. Xu W, Luo G, Meng W, et al. MRAgent: an LLM-based automated agent for causal knowledge discovery in disease via Mendelian randomization. Brief Bioinform. 2025;26(2). doi:10.1093/bib/bbaf140
34. Qu Y, Huang K, Yin M, et al. CRISPR-GPT for agentic automation of gene-editing experiments. Nat Biomed Eng. Published online July 30, 2025. doi:10.1038/s41551-025-01463-z
35. Mehandru N, Hall AK, Melnichenko O, et al. BioAgents: Bridging the gap in bioinformatics analysis with multi-agent systems. Sci Rep. 2025;15(1):39036. doi:10.1038/s41598-025-25919-z
36. Yang EW, Velazquez-Villarreal E. AI-HOPE: an AI-driven conversational agent for enhanced clinical and genomic data integration in precision medicine research. Bioinformatics. 2025;41(7). doi:10.1093/bioinformatics/btaf359
37. Qu Y, Huang K, Yin M, et al. CRISPR-GPT for agentic automation of gene-editing experiments. Nat Biomed Eng. Published online July 30, 2025. doi:10.1038/s41551-025-01463-z
38. Xie E, Cheng L, Shireman J, et al. CASSIA: a multi-agent large language model for automated and interpretable cell annotation. Nat Commun. 2025;17(1):389. doi:10.1038/s41467-025-67084-x
39. Yang EW, Velazquez-Villarreal E. AI-HOPE: An AI-Driven conversational agent for enhanced clinical and genomic data integration in precision medicine research. Genetic and Genomic Medicine. Preprint posted online December 1, 2024. doi:10.1101/2024.11.27.24318113
40. Yang EW, Waldrup B, Velazquez-Villarreal E. AI-HOPE-TGFbeta: A Conversational AI Agent for Integrative Clinical and Genomic Analysis of TGF-β Pathway Alterations in Colorectal Cancer to Advance Precision Medicine. AI. 2025;6(7):137. doi:10.3390/ai6070137
41. Yang EW, Waldrup B, Velazquez-Villarreal E. AI-HOPE-TP53: A Conversational Artificial Intelligence Agent for Pathway-Centric Analysis of TP53-Driven Molecular Alterations in Early-Onset Colorectal Cancer. Cancers. 2025;17(17):2865. doi:10.3390/cancers17172865

42. Yang EW, Waldrup B, Velazquez-Villarreal E. Artificial Intelligence for Precision Oncology: AI-HOPE-MAPK Uncovers Clinically Actionable MAPK Alterations in Colorectal Cancer. AI in Precision Oncology. 2025;2(5):197-206. doi:10.1177/2993091X251380198
43. Yang EW, Waldrup B, Velazquez-Villarreal E. Conversational AI agent for precision oncology: AI-HOPE-WNT integrates clinical and genomic data to investigate WNT pathway dysregulation in colorectal cancer. Front Artif Intell. 2025;8:1624797. doi:10.3389/frai.2025.1624797
44. Yang EW, Waldrup B, Velazquez-Villarreal E. Decoding the JAK-STAT Axis in Colorectal Cancer with AI-HOPE-JAK-STAT: A Conversational Artificial Intelligence Approach to Clinical–Genomic Integration. Cancers. 2025;17(14):2376. doi:10.3390/cancers17142376
45. Yang EW, Waldrup B, Velazquez-Villarreal E. From Mutation to Prognosis: AI-HOPE-PI3K Enables Artificial Intelligence Agent-Driven Integration of PI3K Pathway Data in Colorectal Cancer Precision Medicine. IJMS. 2025;26(13):6487. doi:10.3390/ijms26136487
46. Yang EW, Waldrup B, Velazquez-Villarreal E. Precision Oncology Through Dialogue: AI-HOPE-RTK-RAS Integrates Clinical and Genomic Insights into RTK-RAS Alterations in Colorectal Cancer. Biomedicines. 2025;13(8):1835. doi:10.3390/biomedicines13081835
47. Yang EW, Waldrup B, Velazquez-Villarreal E. Conversational Artificial Intelligence for Integrating Social Determinants, Genomics, and Clinical Data in Precision Medicine: Development and Implementation Study of the AI-HOPE-PM System. JMIR Bioinform Biotech. 2025;6:e76553-e76553. doi:10.2196/76553
48. Kamoi R, Zhang Y, Zhang N, Han J, Zhang R. When Can LLMs Actually Correct Their Own Mistakes? A Critical Survey of Self-Correction of LLMs. Preprint posted online December 3, 2024. doi:10.1162/tacl_a_00713/125177
49. Huang J, Chen X, Mishra S, et al. Large Language Models Cannot Self-Correct Reasoning Yet. arXiv. Preprint posted online March 14, 2024:arXiv:2310.01798. doi:10.48550/arXiv.2310.01798

Figures

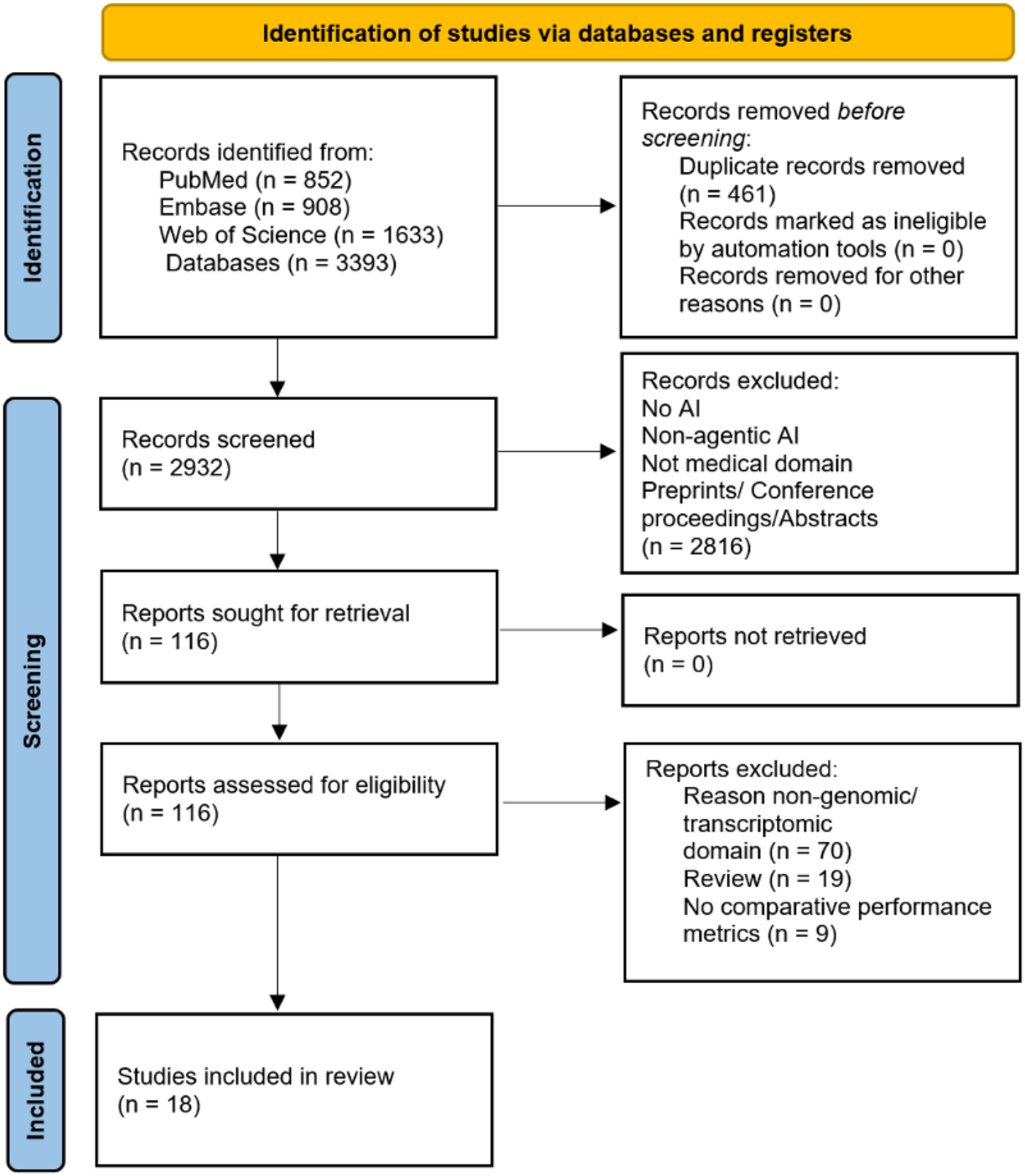

Figure 1

PRISMA Flow Diagram

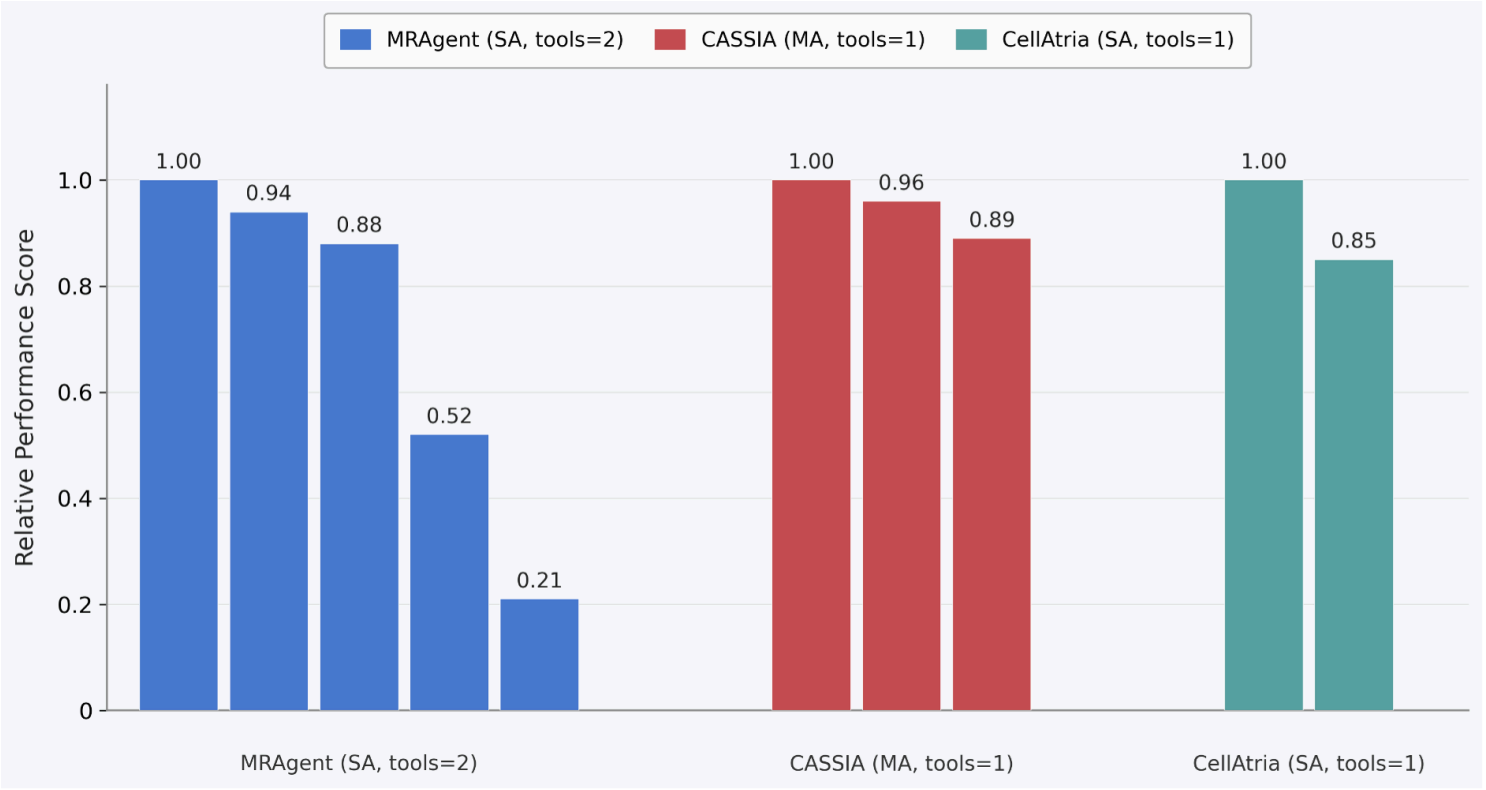

Figure 2

Normalized backbone sensitivity across agentic systems.

Performance was normalized within each system by scaling the best-performing LLM to 1.0. MRAgent (F1, GWAS selection) shows substantial variability (1.00–0.21)[33], CASSIA (accuracy, cell annotation) shows moderate variability (1.00–0.89)[38], and CellAtria (runtime, lower better) shows limited variability (1.00–0.85)[30]. For F1 and accuracy, normalized score = model performance / best performance; for runtime, normalized score = best runtime / model runtime. Values reflect within-system backbone robustness and are not directly comparable across systems. SA = single-agent architecture; MA = multi-agent architecture; tools = number of external tools integrated into the system's workflow.

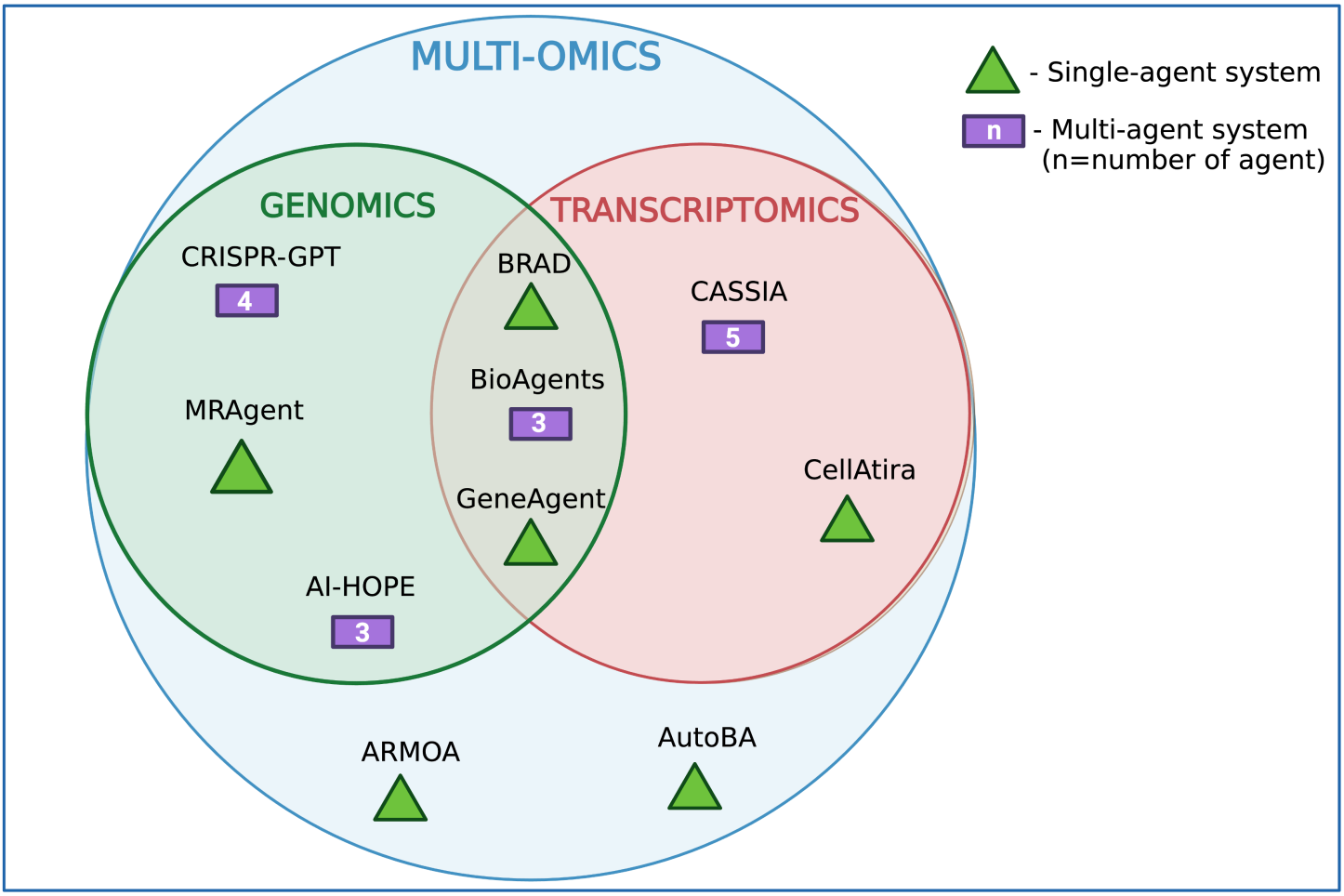

Figure 3

Figure 1. PRISMA Flow Diagram

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendixfinal.pdf