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.

Artificial Intelligence-Driven Self-Healing Bioinformatics Pipelines: A Systematic Review of Automated Failure Detection and Remediation in Omics and Computational Biology Workflows

Bipul Bhattarai
University Of South Dakota https://orcid.org/0009-0001-4810-4007
Sulav Dahal
sulav.dahal@mavs.uta.edu
University of Texas at Arlington https://orcid.org/0009-0008-5736-5293
Naina Maharjan
naina.maharjan@coyotes.usd.edu
University Of South Dakota https://orcid.org/0009-0002-5302-3209

Systematic Review

Keywords: bioinformatics pipelines, self-healing systems, large language models, multiagent orchestration, anomaly detection workflow reproducibility, AIOps, systematic review, omics computing, retrieval-augmented generation

Posted Date: May 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9619074/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.

Artificial Intelligence-Driven Self-Healing Bioinformatics Pipelines: A Systematic Review of Automated Failure Detection and Remediation in Omics and Computational Biology Workflows

Bipul Bhattarai[1*], Naina Maharjan[1], Sulav Dahal[2]

[1]Department of Biomedical Engineering, University of South Dakota, Vermillion, SD 57069, United States of America
[2]Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, United States of America

*Corresponding author. E-mail: bipulbhattarai70@gmail.com
ORCID: 0009-0001-4810-4007

Abstract

Background: Bioinformatics pipelines spanning genomics, transcriptomics, proteomics, and metagenomics face a pervasive reproducibility crisis driven by software dependency drift, resource exhaustion, and non-deterministic tool behaviour. Despite substantial investment in workflow management systems such as Nextflow, Snakemake, and Galaxy, pipeline failure responses remain predominantly manual. Advances in large language models (LLMs), retrieval-augmented generation (RAG), and multi-agent orchestration offer technically credible pathways to automated pipeline failure detection and autonomous remediation, yet their systematic application to bioinformatics-specific infrastructure has not been reviewed.

Results: This PRISMA 2020-compliant systematic review (PROSPERO: CRD420261361756) synthesised 26 studies from six databases covering January 2019 to April 2026, addressing five research questions spanning LLM-based log parsing, multi-agent workflow orchestration, human-in-the-loop governance, fault-tolerant infrastructure patterns, and benchmarking gaps. LLM-RAG frameworks achieved up to 80% workflow step recall and reduced manual curation time by over 90%. Multi-agent systems including BioMaster and MARWA demonstrated superior error recovery across 18 omics modalities and 102 bioinformatics tools, consistently outperforming single-agent baselines. Infrastructure foundations are technically mature, but no included study demonstrated an end-to-end integrated self-healing pipeline combining monitoring, anomaly detection, remediation, and governance-compliant audit trail generation. Evidence quality was moderate (mean 6.2/10; range 5–9).

Conclusions: Governance frameworks, bioinformatics-specific benchmarks, and regulatory alignment for clinical contexts remain critically absent and represent the field's primary bottleneck. We propose a four-layer conceptual model (Infrastructure, Observability, Orchestration, Governance) to organise current evidence and identify research priorities. Three directions warrant priority investment: standardised failure-injection benchmarks, end-to-end integrated pipeline validation in production bioinformatics environments, and LLM fine-tuning on bioinformatics-specific log formats.

Keywords: bioinformatics pipelines; self-healing systems; large language models; multi-agent orchestration; anomaly detection; workflow reproducibility; AIOps; systematic review; omics computing; retrieval-augmented generation

Background

Computational biology has undergone a transformation in scale and complexity over the past decade. The proliferation of high-throughput sequencing technologies — including whole-genome sequencing, single-cell RNA sequencing, long-read sequencing, and spatial transcriptomics — has generated omics datasets of unprecedented volume, demanding multi-step computational pipelines that orchestrate dozens of specialised tools across distributed computing environments [1]. The diversity of omics domains — genomics, transcriptomics, proteomics, metagenomics, and their multi-omics integrations — has further expanded the pipeline landscape, with each domain requiring bespoke analysis chains, software dependencies, and reference databases that evolve continuously.

This complexity has given rise to a pervasive reproducibility crisis. A systematic evaluation of machine learning studies in the life sciences found that fewer than 33% of published results were reproducible, with fewer than 5% of researchers sharing the source code necessary for independent replication [2]. In computational biology specifically, reproducibility failures arise from multiple interacting sources: software dependency drift, in which tool versions diverge between original and replication environments; non-deterministic tool behaviour producing inconsistent outputs across runs; resource contention on shared computing infrastructure; data corruption or silent format errors during pipeline execution; and inadequate provenance tracking that prevents auditing of analytical decisions [3]. These failures are not merely methodological inconveniences: in clinical genomics, where variant calling pipelines inform diagnostic and treatment decisions, irreproducible or silently failing computations carry direct patient safety implications.

Bioinformatics pipeline failures — defined as unintended interruptions, errors, or deviations from expected execution behaviour — represent a primary mechanism through which irreproducibility enters computational biology. Unlike the analytical choices that have received extensive attention in the reproducibility literature, pipeline failures are often silent: a misconfigured alignment step may complete without error messages while producing subtly incorrect output that propagates undetected through downstream variant calling, differential expression analysis, or clinical report generation [4].

Modern workflow management systems — including Nextflow, Snakemake, Galaxy, Cromwell, and the Common Workflow Language (CWL) ecosystem — have substantially improved pipeline reliability through automatic checkpointing, dependency tracking, and environment containerisation [5]. However, these systems address the execution environment rather than the content of failures: they can resume a pipeline from the last successful checkpoint but cannot diagnose why a step failed, assess whether the failure is systematic or transient, propose an appropriate corrective action, or implement that action autonomously while maintaining a compliance-grade audit trail. The response to pipeline failures therefore remains predominantly manual, consuming hours or days and constituting itself a source of undocumented, non-reproducible interventions.

The software engineering community has developed AI-driven self-healing systems — infrastructure that automatically detects failures, diagnoses root causes, and implements corrective actions — demonstrating meaningful reductions in Mean Time to Recovery (MTTR) in general IT contexts [6]. Recent advances in LLM-based log parsing [7], multi-agent orchestration frameworks [8], and RAG have substantially expanded the feasibility of these approaches. RAG is particularly relevant to bioinformatics, where precise tool parameterisation and version-specific command syntax require grounding in domain-specific knowledge bases to avoid the hallucination risk that propagates silently through downstream analyses [9]. These capabilities are directly applicable to bioinformatics pipeline failure modes, yet their systematic evaluation in omics and biomedical computing contexts has not previously been reviewed.

Existing systematic reviews addressing AI in bioinformatics have focused on the application of machine learning to biological data analysis — variant effect prediction, protein structure modelling, single-cell classification — rather than to the computational infrastructure that executes those analyses. Reviews of self-healing and AIOps systems have been conducted within the software engineering and IT operations literature, without bioinformatics-specific evidence, terminology, or regulatory context. The specific intersection — AI-driven self-healing applied to the pipeline infrastructure of omics and biomedical computing — has not been the subject of a prior systematic review.

This systematic review provides the first PRISMA 2020-compliant synthesis of evidence on AI-driven self-healing approaches applied to bioinformatics and computational biology pipeline infrastructure. It makes four primary contributions: a taxonomy of AI approaches as applied to omics pipeline contexts; characterisation of the bioinformatics-specific multi-agent systems that have emerged since 2024; systematic mapping of critical benchmarking, standardisation, and regulatory gaps; and a four-layer conceptual model of AI-driven self-healing bioinformatics infrastructure to guide future research.

Methods

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement [10] and prospectively registered on PROSPERO (CRD420261361756) prior to data extraction.

Review objective and research questions

Five pre-specified research questions (RQs) guided the review: RQ1 — what AI and LLM-based approaches have been applied to automated failure detection and log parsing in bioinformatics pipelines, and what performance metrics have been reported? RQ2 — how have multi-agent orchestration frameworks been applied for managing distributed omics data processing workflows? RQ3 — what governance mechanisms, confidence thresholds, and human-in-the-loop protocols have been proposed or evaluated for AI-driven autonomous remediation in clinical bioinformatics? RQ4 — what infrastructure patterns support self-healing and reproducible bioinformatics pipeline execution? RQ5 — what are the critical gaps in benchmarking, standardisation, and regulatory alignment for AI-driven self-healing systems?

Eligibility criteria

Studies were included if they evaluated AI, LLM, or multi-agent interventions applied to pipeline monitoring, failure detection, log parsing, or automated remediation in bioinformatics or scientific computing environments, published from January 2019 to April 2026. Studies addressing only biological data analysis without pipeline infrastructure focus were excluded, as were opinion pieces, editorials, and theoretical papers without empirical evaluation. A full PICOS eligibility table is provided as Additional file 1.

Search strategy

Systematic searches were conducted in six databases: PubMed, ACM Digital Library, IEEE Xplore, arXiv, bioRxiv, and Google Scholar, executed in April 2026. Search strings combined terms across three conceptual domains: bioinformatics pipeline contexts; AI, LLM, and multi-agent intervention terms; and self-healing, fault detection, and automation outcome terms. Forward citation searching was conducted on all included studies. Complete search strings are provided as Additional file 5.

Study selection and screening

Screening proceeded in two stages: title and abstract screening followed by full-text assessment. Both stages were conducted by the primary reviewer (B.B.), with a 20% random sample ($n = 19$) independently re-screened after a minimum two-week interval by a second team member (N.M.) to assess intra-rater consistency. Raw percentage agreement was 84.2% (16/19 records). Cohen's kappa was $\kappa = -0.08$, which is unreliable as a reliability indicator in this context because the heavily skewed distribution of screening decisions — where the vast majority of records were excluded — causes expected agreement to approach observed agreement, artificially deflating kappa toward zero or negative values [11]. Percentage agreement is therefore reported as the primary intra-rater reliability metric, consistent with recommended practice for systematic reviews in which one screening decision predominates [11]. All screening decisions are documented in Additional file 2.

Data extraction

Data were extracted using a pre-specified extraction form including: bibliographic identifiers, study design, bioinformatics domain, AI method, workflow manager or computing infrastructure, deployment context, primary outcome metrics, and study limitations. Data extraction is documented in Additional file 3.

Quality and risk of bias assessment

Methodological quality was assessed using an adapted Mixed Methods Appraisal Tool (MMAT) [12], supplemented with four additional criteria for AI system evaluation: availability of evaluation datasets; independence of evaluation metrics from training data; presence of a formal comparator; and generalisability beyond the specific implementation. Each study was assigned a quality score of 0–10, with scores of 7–10 classified as high quality, 5–6 as moderate, and below 5 as low. Studies were not excluded on quality grounds alone. Full quality assessment scores are provided in Additional file 4.

Data synthesis

Narrative synthesis was adopted given substantial heterogeneity in study designs, outcome metrics, and deployment contexts, which precluded quantitative meta-analysis. Findings were organised thematically

by research question and AI approach. Certainty of evidence was assessed using an adapted GRADE approach [13] (high, moderate, low, or very low) applied at the level of each synthesised finding.

PRISMA compliance and protocol deviations

The review adhered to PRISMA 2020 guidelines throughout; the completed checklist is provided as Additional file 6. One protocol deviation occurred: scope was broadened to include general scientific computing studies alongside bioinformatics-specific evidence, as pilot searching confirmed insufficient bioinformatics-specific literature to support a domain-only review. This deviation is documented in the PROSPERO record (version 1.1, updated 7 April 2026).

Results

Study selection and characteristics

The systematic search yielded 152 records. Following deduplication ($n = 28$ removed), 124 records were screened at title and abstract stage. Sixty-eight records were excluded: wrong domain or focus ($n = 48$); wrong intervention type ($n = 20$). Fifty-six records were retrieved for full-text assessment, of which 30 were excluded: conference proceedings with no identifiable in-scope paper ($n = 18$), wrong deployment context ($n = 5$), late-stage duplicates ($n = 5$), no empirical evaluation ($n = 2$). The final corpus comprised 26 studies (Fig 1).

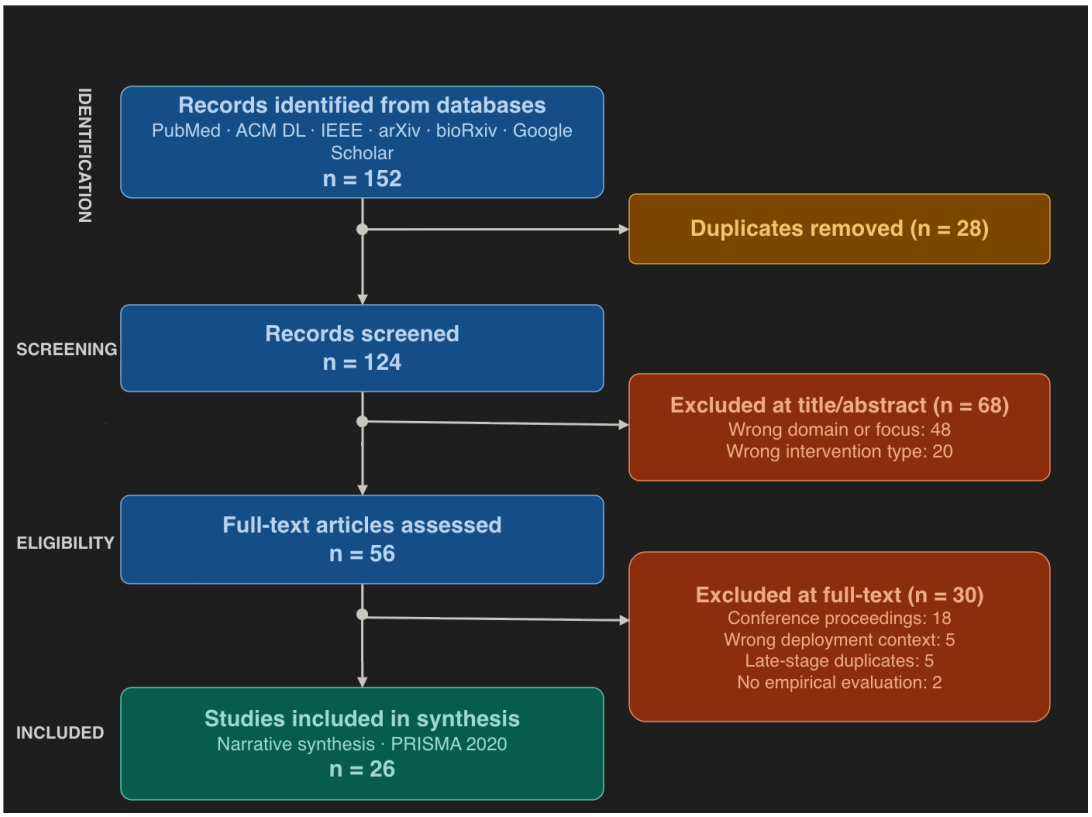


Fig 1: **PRISMA 2020 flow diagram showing study selection.** Records were identified from six databases (PubMed, ACM Digital Library, IEEE Xplore, arXiv, bioRxiv, Google Scholar) and screened through two stages. Exclusion reasons are shown for both screening stages.

Studies spanned scientific HPC environments ($n = 9$), general data pipeline systems relevant to bioinformatics ($n = 8$), bioinformatics-specific environments ($n = 6$), and healthcare IT infrastructure ($n = 3$). Fourteen studies (54%) reported quantitative outcomes; seven (27%) were proof-of-concept implementations; and five (19%) were survey or framework papers. Sixty-five percent of studies ($n = 17$) were published in 2025. Mean quality score was 6.2/10 (range 5–9). Table 1 summarises the 26 included studies.

Table 1: **Summary of 26 included studies.** Abbreviations: Gen. bio = General bioinformatics; Sci. HPC = Scientific HPC; Gen. pipeline = General data pipeline; HC IT = Healthcare IT; PoC = proof of concept; CS = case study; Sur = survey; Emp = empirical; Q = quality score /10.

ID	First Author (Year)	Domain	AI Approach	Design	RQ(s)	Q
1	Kapoor & Narayanan (2023)	Gen. bio	Reproducibility benchmarking framework	Sur	RQ5	8
2	Anon. (2022)	Sci. HPC	Staging-based workflow data resiliency	Emp	RQ4	6
3	Anon. (2022)	Sci. HPC	Kubernetes scheduling taxonomy	Sur	RQ4	6
4	Anon. (2023)	Sci. HPC	Scalable composable workflow architecture	Emp	RQ4	6
5	Anon. (2024)	Gen. bio	LLM-based workflow model generation	PoC	RQ1	5
6	Anon. (2025)	Sci. HPC	Secure API-driven research automation	CS	RQ4	6
7	Rosendo et al. (2025)	Sci. HPC/ORNL	Multi-agent LLM orchestration for HPC	PoC	RQ2	7
8	Shin et al. (2025)	Sci. HPC	State machine agentic workflow framework	Sur	RQ2, RQ5	8
9	Nellutla et al. (2025)	Gen. pipeline	Serverless autonomous ML pipeline	PoC	RQ4	5
10	Verma et al. (2025)	HC IT	AI observability: Prometheus + BigPanda	Emp	RQ1, RQ3	6
11	Anon. (2025)	Gen. pipeline	Resilient intelligent ETL architecture	PoC	RQ4	5
12	Carnevale et al. (2025)	Sci. HPC	Bi-LSTM + federated learning for HPC	Emp	RQ1	7
13	Gwinn et al. (2025)	Genomics/HPC	Error analysis: drug response pipeline	CS	RQ1, RQ4	7
14	Satyanarayanan (2022)	Gen. pipeline	Self-healing framework for AI pipelines	PoC	RQ3, RQ4	5
15	Bradfield et al. (2025)	Gen. pipeline	GenAI for automated pipeline repair	PoC	RQ1, RQ3	5
16	Chakraborty (2025)	Gen. pipeline	AI agents for self-healing ETL	PoC	RQ2, RQ3	5
17	Hughes & Luke (2024)	Gen. pipeline	Self-healing pipeline impact on AI fairness	Emp	RQ3, RQ5	5
18	Anand & Sharma (2024)	HC IT	Self-healing pipelines for Medicaid data	CS	RQ1, RQ3	6
19	Dovzhenko et al. (2020)	Gen. bio	AI-augmented bioinformatics monitoring	PoC	RQ1	5
20	Bhaskar et al. (2025)	HC IT	Predictive server failure detection	CS	RQ1, RQ4	5
21	Alam et al. (2026)	Gen. bio	Evolutionary intrusion detection system	Emp	RQ1	6
22	Wang & Wang (2025)	Gen. bio	BioWorkflow: LLM-RAG workflow extraction	Emp	RQ1, RQ5	9
23	Hou et al. (2025)	Multi-omics	MARWA: multi-agent RAG automation	Emp	RQ2	7
24	Gu et al. (2025)	Gen. bio	Survey: LLMs for ML workflow construction	Sur	RQ1, RQ5	7
25	Alam & Roy (2025)	Gen. bio	LLMs for scientific workflow development	Emp	RQ1, RQ2	7
26	Su et al. (2025)	Multi-omics	BioMaster: multi-agent bioinformatics	Emp	RQ2	8

RQ1: LLM-based and AI-driven failure detection and log parsing

LLM-based workflow extraction and reconstruction

The strongest empirical contribution to RQ1 was BioWorkflow [14], a peer-reviewed LLM-RAG framework published in *Briefings in Bioinformatics*. Applied to 100 expert-annotated bioinformatics publications, BioWorkflow achieved approximately 80% recall of workflow steps, compared with approximately 20% for the BCO-RAG baseline. Reproducibility, completeness, and accuracy each improved by more than 20% relative to GPT-o4-mini, and manual curation time was reduced from over one hour to three to five minutes per paper. Complementary evidence from Alam and Roy [15] demonstrated LLM generation of Nextflow-compatible pipelines with automatic provenance logging, while Gu et al. [16] identified a near-complete absence of bioinformatics-specific evaluation datasets for LLM-based workflow construction as a critical gap.

ML-based anomaly detection and resource prediction

Carnevale et al. [17] presented a federated Bi-LSTM approach for CPU utilisation prediction in distributed HPC resource management, directly applicable to multi-site genomics consortia requiring cross-institutional model training under privacy constraints. Gwinn et al. [18] applied statistical standardisation and distribution analysis to error detection in a drug response prediction workflow deployed on the Argonne Leadership Computing Facility supercomputer. Verma et al. [19] reported an AI observability implementation integrating Prometheus with BigPanda alert correlation, achieving approximately 40% reduction in MTTR and 75% reduction in alert noise.

Bioinformatics-specific AI monitoring frameworks

Dovzhenko et al. [20] presented the earliest bioinformatics-specific contribution (2020), describing an AI-augmented shell-based monitoring framework for omics pipelines integrating anomaly detection, resource forecasting, and self-healing capabilities. Alam et al. [21] contributed an evolutionary algorithm-inspired intrusion detection system evaluated on log traffic mimicking GA4GH and NIH SRA pipeline architectures. Anand and Sharma [22] reported a self-healing pipeline implementation for healthcare data processing combining anomaly detection and compliance-aware governance. Across the 13 RQ1 studies, heterogeneity in outcome metrics was substantial, precluding formal meta-analysis.

RQ2: Multi-agent orchestration frameworks

BioMaster [23] presented the most comprehensively evaluated multi-agent bioinformatics system in the corpus, employing a dual RAG system alongside a dedicated Debug Agent for real-time error detection and correction. Across 49 tasks spanning 18 omics modalities and 102 bioinformatics tools, BioMaster substantially outperformed existing automated systems. MARWA [24] introduced contrastive learning-enhanced tool retrieval, step-by-step generation with per-stage error handling, and a two-stage evaluation scheme on 40 curated and 2,270 large-scale tasks, consistently outperforming single-agent baselines. A consistent pattern across the three bioinformatics-specific multi-agent systems (Studies 23, 25, 26) was the central role of RAG, suggesting it represents an emerging architectural standard.

Rosendo et al. [25] described a modular multi-agent architecture for autonomous cross-facility experimentation at ORNL, integrating a W3C PROV-compliant provenance infrastructure using the Flowcept system. Shin et al. [26] provided the theoretical foundation for this field, proposing a state machine abstraction framework spanning from static DAG workflows to intelligent multi-agent swarms, projecting 10- to 100-fold acceleration in scientific discovery as the potential outcome.

RQ3: Governance mechanisms and human-in-the-loop protocols

Seven studies addressed governance aspects. Satyanarayanan [27] proposed a foundational governance framework delineating automated fault detection, confidence-based escalation, and audit trail generation layers. Verma et al. [19] demonstrated confidence-based human-in-the-loop (HITL) architecture in healthcare IT. Hughes and Luke [28] identified a governance dimension not considered by other studies: inconsistent pipeline self-healing propagates systematic biases into downstream AI model fairness metrics, with potential health equity implications for clinical genomics.

Four critical governance gaps were identified: absence of confidence threshold calibration for bioinformatics contexts; no regulatory alignment with GA4GH, FAIR, HIPAA, or ISO 15189 frameworks; audit trail implementations described architecturally but not empirically validated; and no framework addressing the speed-versus-verification tradeoff in time-sensitive clinical genomics remediation.

RQ4: Infrastructure patterns

Nine studies addressed infrastructure patterns. The Kubernetes scheduling taxonomy identified ML-based scheduling as an emerging capability for scientific workloads, directly relevant to containerised bioinformatics pipelines executing through Nextflow, Snakemake, and Cromwell. Nellutla et al. [29] described a fully autonomous ML pipeline architecture using serverless cloud functions, aligning with cloud-native genomics platforms such as AWS HealthOmics and Google Cloud Life Sciences. Data resiliency mechanisms established an empirical basis for staging-layer fault tolerance in HPC scientific workflows critical for long-running genomics analyses.

RQ5: Critical gaps in benchmarking, standardisation, and regulatory alignment

Kapoor and Narayanan [2] documented that fewer than 33% of ML studies in the life sciences are reproducible, identifying standardised evaluation frameworks as primary mechanisms required. Gu et al. [16] identified a near-complete absence of bioinformatics-specific evaluation datasets for LLM-based workflow construction. Shin et al. [26] identified the absence of standardised benchmarks for multi-step agentic workflow evaluation as the single most important impediment to progress. Hughes and Luke [28] identified a gap at the intersection of pipeline reliability and AI fairness. No included study addressed alignment between AI-driven self-healing and regulatory frameworks governing clinical bioinformatics, including GA4GH standards, FAIR principles, FDA Software as a Medical Device guidance, GDPR and HIPAA data governance, or ISO/IEC quality management standards.

Evidence quality assessment

Quality assessment revealed moderate to low overall evidence quality (mean 6.2/10, range 5–9). Nine studies (35%) scored 7 or above (high quality); 12 (46%) scored 5–6 (moderate); and 5 (19%) scored below 5 (low). The most common quality deficiencies were: absence of a formal comparator (58%), non-public evaluation datasets (62%), and single-environment evaluation without cross-domain validation (69%).

Of 26 studies, 24 (92%) reported positive findings, consistent with publication bias in an emerging proof-of-concept-dominated field.

Sixty-five percent of studies ($n = 17$) were published in 2025 alone, with infrastructure-level contributions dominating 2022–2023 and LLM-based and multi-agent systems emerging as distinct trajectories from 2024 onward.

Discussion

This systematic review synthesised 26 studies addressing AI-driven automated failure detection, multi-agent orchestration, governance, and infrastructure patterns for self-healing bioinformatics pipelines. The principal finding is that the field is in rapid but uneven emergence: LLM-RAG and multi-agent approaches demonstrate genuine capability for automated workflow construction, error recovery, and pipeline monitoring, yet rigorous evaluation in bioinformatics-specific contexts remains strikingly scarce.

Interpretation of findings by research question

The convergence of three independent studies [14, 15, 24] on RAG as the architectural foundation for reliable bioinformatics workflow automation is theoretically significant: RAG addresses the fundamental challenge of providing precise, tool-specific knowledge that general pre-training cannot reliably supply, directly reducing the hallucination risk that is particularly consequential where silent errors propagate through downstream analyses.

BioMaster [23] and MARWA [24] represent the current state of the art, employing role-specialised agents, dual RAG retrieval, and dedicated error recovery. Using the state machine framework of Shin et al. [26], current bioinformatics automation occupies the Static-Pipeline quadrant; BioMaster and MARWA represent progress toward Adaptive-Pipeline. The fully autonomous Intelligent-Swarm quadrant remains unrealised in bioinformatics.

A significant asymmetry exists between the sophistication of detection and remediation approaches and the maturity of governance frameworks. The finding of Hughes and Luke [28] — that self-healing mechanisms have downstream consequences for AI model fairness — introduces a governance dimension not previously considered in the AIOps literature. In bioinformatics, inconsistent autonomous remediation could introduce systematic biases into clinical risk scores or pharmacogenomics reports, with potential health equity implications.

The technical foundations for self-healing bioinformatics pipelines are well-established, but no included study demonstrated an end-to-end integrated self-healing bioinformatics pipeline combining all four layers: real-time monitoring, AI-driven anomaly detection, multi-agent remediation, and governance-compliant audit trail generation.

Bioinformatics-specific benchmarking for AI-driven pipeline automation is effectively absent. This review itself is constrained by this gap: heterogeneity of outcome metrics precluded formal meta-analysis and required narrative synthesis.

Four-layer conceptual model

Integrating evidence across all five RQs, this review proposes a four-layer conceptual model of AI-driven self-healing bioinformatics infrastructure (Fig 2). Layer 1 (Infrastructure: containerised, fault-tolerant execution) is well-evidenced and technically mature. Layer 2 (Observability: LLM log parsing, ML anomaly detection, resource prediction) has strong DevOps evidence with limited bioinformatics-specific evaluation. Layer 3 (Orchestration: multi-agent remediation, RAG-grounded tool coordination) shows rapid 2025 progress without production bioinformatics validation. Layer 4 (Governance: confidence-based HITL, audit trails, regulatory compliance) is the least mature, with only foundational frameworks and no bioinformatics-specific calibration or regulatory alignment. The field's primary bottleneck is not technical capability at Layers 1–3 but governance maturity at Layer 4 and the absence of end-to-end integrated implementations.

Positioning relative to existing literature

No prior systematic review has synthesised evidence at the intersection of AIOps and bioinformatics pipeline infrastructure. The three similar PROSPERO registrations identified at protocol registration

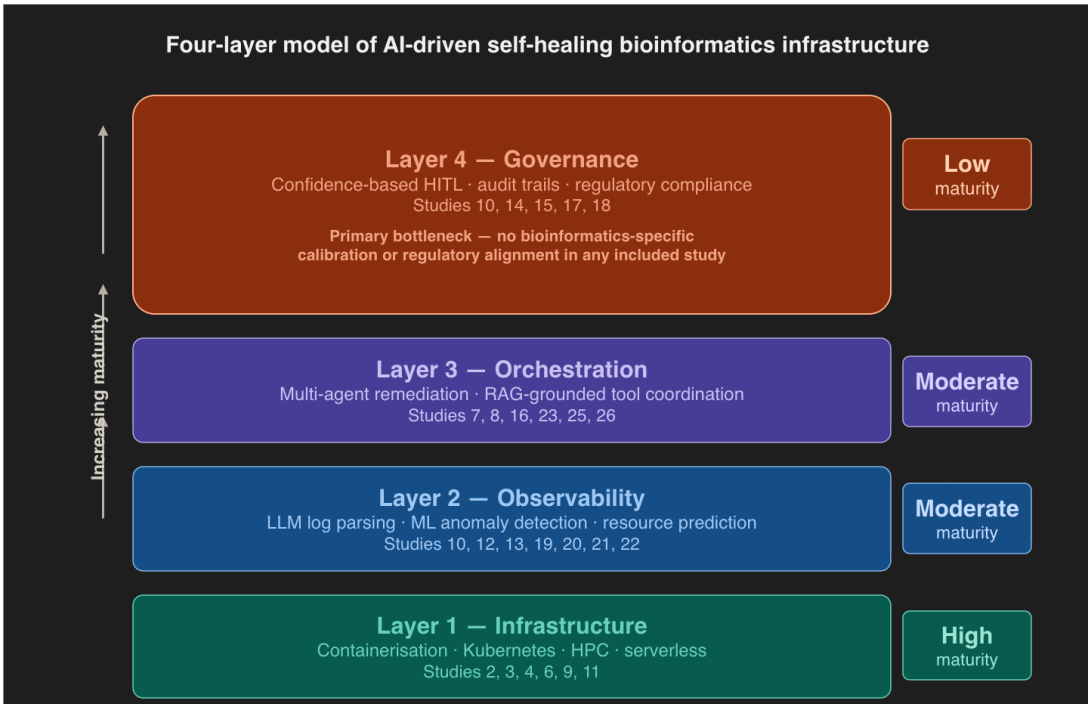


Fig 2: **Four-layer conceptual model of AI-driven self-healing bioinformatics infrastructure.** The four layers progress from technically mature (Layer 1, Infrastructure) through moderate evidence (Layers 2–3, Observability and Orchestration) to critically underdeveloped (Layer 4, Governance). The field's primary bottleneck is at Layer 4.

each addressed adjacent biological discovery, NLP, or clinical AI evidence methodology questions — distinct from this review's focus on the computational systems that manage, monitor, and recover workflow execution.

Limitations

This review has several limitations. First, the narrow evidence base — 26 studies, 65% published in 2025 — reflects an emerging field rather than a mature literature, and limits generalisability of conclusions. Second, the single primary reviewer design, while supplemented by 20% independent re-screening, introduces the possibility of selection bias. Third, heterogeneity of outcome metrics across studies precluded quantitative synthesis and required reliance on narrative methods. Fourth, the scope broadening to include general scientific computing studies, documented as a protocol deviation, may have introduced evidence from contexts only partially analogous to bioinformatics deployments.

Conclusions

This systematic review provides the first PRISMA 2020-compliant synthesis of evidence on AI-driven self-healing approaches applied to bioinformatics and computational biology pipeline infrastructure, synthesising 26 studies across five research questions. LLM-RAG architectures demonstrate genuine capability for automated bioinformatics workflow extraction, with BioWorkflow achieving 80% workflow step recall and reducing manual curation time by over 90%. Multi-agent systems including BioMaster and MARWA represent a qualitative advance over single-agent approaches, incorporating real-time error recovery across diverse omics modalities. Infrastructure foundations are technically mature, but end-to-end integration of observability, orchestration, and governance layers within a production bioinformatics deployment has not yet been demonstrated. Governance frameworks, confidence threshold calibration, and regulatory alignment for clinical bioinformatics represent the field's most critical unaddressed gap.

Three research directions warrant priority investment. First, the field urgently requires standardised bioinformatics-specific benchmarks for AI-driven pipeline failure detection and remediation. Second, an end-to-end integrated self-healing pipeline demonstration in a production bioinformatics environment

would validate the four-layer conceptual model and establish a reference implementation. Third, fine-tuning open-source LLMs on bioinformatics-specific log formats — Nextflow stdout, Snakemake error logs, SLURM scheduler messages, and container exit codes — would address the core limitation that current approaches rely on general-purpose models.

Declarations

Ethics approval and consent to participate

Not applicable. This systematic review did not involve human participants, human data, animal subjects, or biological tissue. No ethics approval was required.

Consent for publication

Not applicable. This manuscript does not contain data from any individual person.

Availability of data and materials

This systematic review did not generate primary data or analysis code. All data underlying the findings are available within the manuscript and its supplementary files. The complete screening log for all 152 identified records is provided as Additional file 2. Full data extraction for all 26 included studies is provided as Additional file 3. Quality assessment scores are provided as Additional file 4. Complete database search strings are provided as Additional file 5. The completed PRISMA 2020 checklist is provided as Additional file 6. The review protocol is publicly registered on PROSPERO (CRD420261361756) and accessible at: `https://www.crd.york.ac.uk/PROSPERO/view/CRD420261361756`

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for this study. The authors conducted this work independently with no financial support from any agency in the public, commercial, or not-for-profit sectors.

Authors' contributions

B.B. conceived the study, developed the review protocol, conducted the systematic database search, performed title/abstract screening, full-text assessment, data extraction, quality assessment, and narrative synthesis, and drafted the manuscript. N.M. contributed to methodology development, independently re-screened 20% of records to assess intra-rater consistency, contributed to data extraction, and critically reviewed the manuscript. S.D. contributed to formal analysis, data extraction verification, and critically reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Claude (Anthropic, version Sonnet 4.6) was used to assist with structuring and drafting sections of this manuscript. All content was independently reviewed, verified against source materials, and revised by the authors. Intellectual responsibility for all claims, interpretations, and conclusions rests entirely with the authors. Claude is not listed as an author. The authors thank the research computing communities at the University of South Dakota and the University of Texas at Arlington for institutional support.

Authors' information

B.B. is a researcher at the University of South Dakota with interests in biomedical informatics, computational biology, and AI-driven automation. N.M. is a researcher at the University of South Dakota working in biomedical informatics and applied machine learning. S.D. is a researcher at the University of Texas at Arlington with expertise in computer science, artificial intelligence, and machine learning.

Additional files

Additional file 1. Full PICOS eligibility criteria table. Complete inclusion and exclusion criteria across Population, Intervention, Comparator, Outcomes, Study design, Date range, and Language dimensions.

Additional file 2. Screening log. Complete record of all 152 identified records with title, source, screening decision at each stage, and reason for exclusion where applicable.

Additional file 3. Data extraction workbook. Full data extraction for all 26 included studies including bibliographic details, AI method, deployment context, outcome metrics, and quality assessment scores.

Additional file 4. Quality assessment scores. Adapted MMAT scores for all 26 included studies with item-level ratings and overall classification.

Additional file 5. Complete search strings. Database-specific search strings for PubMed, ACM Digital Library, IEEE Xplore, arXiv, bioRxiv, and Google Scholar with date of execution and number of results returned.

Additional file 6. PRISMA 2020 checklist. Completed 27-item checklist confirming reporting compliance.

References

1. Wratten L, Wilm A, Göke J. Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers. Nat Methods. 2021;18(10):1161-8. doi:10.1038/s41592-021-01254-9.

2. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. doi:10.1016/j.patter.2023.100804.

3. Djaffardjy M, Marchment G, Suter F, et al. Developing and reusing bioinformatics data analysis pipelines using scientific workflow systems. Comput Struct Biotechnol J. 2023;21:2075-85. doi:10.1016/j.csbj.2023.03.003.

4. Grüning BA, Sallou O, Moreno P, et al. Recommendations for the packaging and containerizing of bioinformatics software. F1000Res. 2019;7:742. doi:10.12688/f1000research.15140.2.

5. Di Tommaso P, Chatzou M, Floden EW, Prieto Barja P, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35(4):316-9. doi:10.1038/nbt.3820.

6. Dang Y, Lin Q, Huang P. AIOps: real-world challenges and research innovations. IEEE Softw. 2019;36(4):7-11. doi:10.1109/MS.2019.2908440.

7. Chen Z, Liu S, Huang J, et al. LogParser-LLM: advancing efficient log parsing with large language models. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York: ACM; 2024. ArXiv:2408.13727.

8. Liu H, Zhang Y, Zhao X, et al.. Multi-agent collaboration mechanisms: a survey of LLMs; 2025. ArXiv:2501.06322 [Preprint]. Available from: https://arxiv.org/abs/2501.06322.

9. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems. vol. 33; 2020. p. 9459-74.

10. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71. doi:10.1136/bmj.n71.

11. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol. 1990;43(6):543-9. doi:10.1016/0895-4356(90)90159-M.

12. Hong QN, Fàbregues S, Bartlett G, et al. The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. Educ Inf. 2018;34(4):285-91. doi:10.3233/EFI-180221.

13. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336:924-6. doi:10.1136/bmj.39489.470347.AD.

14. Wang Y, Wang J. BioWorkflow: retrieving comprehensive bioinformatics workflows from publications. Brief Bioinform. 2025;26(6):bbaf571. doi:10.1093/bib/bbaf571.

15. Alam S, Roy S. From prompt to pipeline: large language models for scientific workflow development in bioinformatics; 2025. ArXiv Preprint.

16. Gu J, Chen Z, Li X, et al. Large language models for constructing and optimizing machine learning workflows: a survey. ACM Trans Softw Eng Methodol. 2025. doi:10.1145/3712003.

17. Carnevale A, Sallou O, Hosseini M, et al. Private distributed resource management: predicting CPU utilization using Bi-LSTM and federated learning. In: Proceedings of IEEE CCGrid 2025. IEEE; 2025. .

18. Gwinn A, Rodriguez M, Wozniak J, Stevens R, et al. A workflow for error analysis for drug response prediction pipelines. In: Proceedings of the SC '25 Workshops. IEEE/ACM; 2025. .

19. Verma A, Sharma P, Singh R, et al. AI-driven observability for real-time incident management in enterprise environments. In: Proceedings of IEEE ICRTEECT 2025. IEEE; 2025. .

20. Dovzhenko A, Korol O, Bilous V, et al.. AI-augmented monitoring of bioinformatics pipelines; 2020. Preprint. Available from ResearchGate.

21. Alam S, Rahman M, Islam J, et al. Adaptive evolution-inspired algorithm for intrusion detection in bioinformatics pipeline environments. Inform Health Bioinform. 2026.

22. Anand R, Sharma P. Self-healing data pipelines for handling anomalies in Medicaid and CHIP data. Int J Artif Intell Biomed. 2024.

23. Su H, Long W, Zhang Y, et al.. BioMaster: multi-agent system for automated bioinformatics analysis workflow; 2025. BioRxiv [Preprint]. 2025 [posted 2025 Jan 23; cited 2026 Apr]. Available from: https://www.biorxiv.org/content/10.1101/2025.01.23.634608v1. doi: 10.1101/2025.01.23.634608.

24. Hou Y, Qiu Y, Nguyen VH, et al.. MARWA: multi-agent retrieval-augmented framework for bioinformatics workflow automation; 2025. Preprint. Under review at OpenReview.

25. Rosendo D, Ghosal S, McDonnell M, Miller R, Skluzacek TJ, Haley J, et al. AI agents for enabling autonomous experiments at ORNL's HPC and manufacturing user facilities. In: Proceedings of the SC '25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis (XLOOP 2025). New York: ACM; 2025. doi:10.1145/3731599.3767592.

26. Shin W, Souza R, Rosendo D, Suter F, Wang F, Balaprakash P, et al. The (r)evolution of scientific workflows in the agentic AI era: towards autonomous science. In: Proceedings of the SC '25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis. New York: ACM; 2025. doi:10.1145/3731599.3767580.

27. Satyanarayanan V. Foundational framework for self-healing data pipelines in AI engineering. Int J Artif Intell Data Sci Mach Learn. 2022.

28. Hughes T, Luke S. The impact of self-healing data pipelines on performance and fairness of AI models; 2024. Preprint. Available from ResearchGate.

29. Nellutla SK, Vijaya Lakshmi N, Nagaraju P, et al. NoOps for AI: fully autonomous ML pipelines using serverless cloud functions. In: Proceedings of the IEEE Global Conference on Wireless Communications and Networking (GCWCN 2025). IEEE; 2025. .

Supplementary Files

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

- DataExtractionWorkbookfinal.xlsx
- SystematicReviewScreeningWorkbookfinal.xlsx
- PRISMAChecklistBhattarai2026.pdf