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.

Comparative benchmarking of AI research agents for omics data interpretation

Nikita V. Basov
Westlake University
Yosef K. Tirta
Westlake University
Zihan Li
Westlake University
Huan Zhou
First Affiliated Hospital of Anhui Medical University
Kiryl D. Piatkevich
kiryl.piatkevich@westlake.edu.cn

Westlake University

Research Article

Keywords: AI scientist, AI benchmarking, bioinformatic analysis, proteomics, transcriptomics, metabolomics

Posted Date: August 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10437888/v1

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

Additional Declarations: No competing interests reported.

Comparative benchmarking of AI research agents for omics data interpretation

Nikita V. Basov[1&], Yosef K. Tirta[1,2&], Zihan Li[1,3], Huan Zhou[4,5,6], Kiryl D. Piatkevich[1,7,8*]

[1]School of Life Sciences, Westlake University, Hangzhou, Zhejiang, China

[2]China-ASEAN College of Marine Sciences, Xiamen University Malaysia, Sepang, Selangor 43900, Malaysia

[3]Coulter Department of Biomedical Engineering, Georgia Institute of Technology, GA 30332, USA

[4]Clinical Research Hospital, The First Affiliated Hospital of Anhui Medical University, Hefei, Anhui, 230022, China

[5]Trial Research Center, The First Affiliated Hospital of Bengbu Medical University, Bengbu, Anhui, 233004, China;

[6]Anhui Province Jointly-Constructed Key Laboratory of Innovative Drugs Pharmaceutical Research and Clinical Evaluation, Bengbu Medical University, Bengbu, Anhui, 233030, China.

[7]Westlake Laboratory of Life Sciences and Biomedicine, Westlake University, Hangzhou, Zhejiang, China

[8]Institute of Basic Medical Sciences, Westlake Institute for Advanced Study, Hangzhou, Zhejiang, China

[&]These authors contributed equally

*Correspondence: kiryl.piatkevich@westlake.edu.cn

Keywords: AI scientist; AI benchmarking; bioinformatic analysis; proteomics; transcriptomics; metabolomics;

Abstract
AI research agents, which are claimed to accelerate or even replace human bioinformatics analysis, are being deployed rapidly in life sciences and biomedical research. However, many scientists lack a systematic approach to assessing their utility and to determining which agent is best suited to a given task. Here we quantitatively benchmark three biomedical AI research agents, including K-Dense (Gemini 2.5 Pro), Finch from Edison Scientific, and Biomni (Claude family), against ChatGPT, which our survey identified as the most commonly used approach for analysing omics datasets. For benchmarking, we performed bioinformatic analyses of metabolomic, bulk proteomic and bulk transcriptomic datasets, which are among the most frequently requested workflows submitted to the deployed AI agents. We evaluated all four agents using a process-anchored framework that scores each output on process quality, result reporting, and interpretation depth, while a domain-aware ordinal rubric is applied in parallel. Eight independent runs per agent per modality, together with bootstrap resampling, alternative scoring definitions, output-volume controls, and within-agent reproducibility measurements, quantified uncertainty. No single agent dominated across all modalities. K-Dense led in metabolomics with the most thorough, reproducible reports; Finch produced the best-organized reports and won proteomics and transcriptomics; Biomni offered the strongest biological interpretation but frequently omitted quality control and preprocessing; ChatGPT, without workflow scaffolding, managed only basic statistics and finished last.Furthermore, agent architecture, not base-model capability, dominates what each agent outputs, with process-anchored evaluation reordering the field relative to accuracy alone. Our benchmark and rubric provide the community with an auditable basis for choosing among AI research agents for omics analysis.

We are currently witnessing a rapid acceleration in the generation of scientific data, particularly in the domains of proteomics, transcriptomics, and metabolomics[1,2,3]. The rapid expansion of omics datasets creates a real bottleneck in data analysis and interpretation. Specifically, these datasets promote data-driven rather than hypothesis-driven research. In this mode, biological interpretation relies entirely on extracting new biological insights from the dataset itself, without preconceived hypotheses[4].

In this context, bioinformaticians experience a bottleneck in analysing and interpreting the vast datasets. AI agents that can operate directly on large volumes of pre-processed omics data, e.g., expression tables, can significantly accelerate bioinformatic analysis, either assisting expert bioinformaticians or, in some cases, substituting for parts of the standard analysis workflow[5,8]. In the past two years, several AI agents specifically optimized for multi-omics data analysis have been introduced, including K-Dense[9], Biomni[7], and Finch (from Edison Scientific https://edisonscientific.com/). These agents allow scientists, even without a coding or bioinformatic background, to interact with their datasets and analyze them using standard bioinformatic workflows, and to generate new hypotheses to test, all through a user-friendly interface. Notably, our independent survey of practitioners revealed that many scientists in the field still rely on general-purpose ChatGPT LLMs to analyse omics data tables. Together, this rapid deployment creates a pressing need for careful benchmarking of these agents on standardized datasets[6,10,30]. Here, we perform a systematic benchmarking of the three biomedical agents (K-Dense, Biomni, Finch) against ChatGPT on three standard datasets in metabolomics, transcriptomics, and proteomics — the three most common data types submitted to such agents according to personal communication with Angela Yiu from Edison Scientific. We evaluated each agent using a standardized prompt, followed by a structured analysis of all outputs produced, including text, figures, and tables and ranking performance of each agent for three modalities.

Results and discussion

Benchmark design and analytical framework

This experimental design benchmarked three recently reported biomedical AI research agents and a general-purpose AI agent in standardized omics analysis across three modalities: targeted LC-MS/MS metabolomics, label-free mass spectrometry bulk proteomics, and bulk transcriptomics. The biomedical AI research agents included K-Dense (Gemini 2.5 Pro), Finch by Edison Scientific (specific LLM is not disclosed), and Biomni (Claude LLM-based)[7], while ChatGPT (GPT-5.5 in medium-reasoning mode) was selected as the only general-purpose AI agent due to its popularity. A schematic benchmark of the study is presented in **Fig. 1**. The base LLM model of Finch was not publicly disclosed at the time of the study; thus, Finch was treated throughout as a deployed system rather than a known underlying language model. This design choice mirrors an agent-benchmarking approach to system evaluation, assessing systems as they are deployed rather than their underlying models alone[8–14].

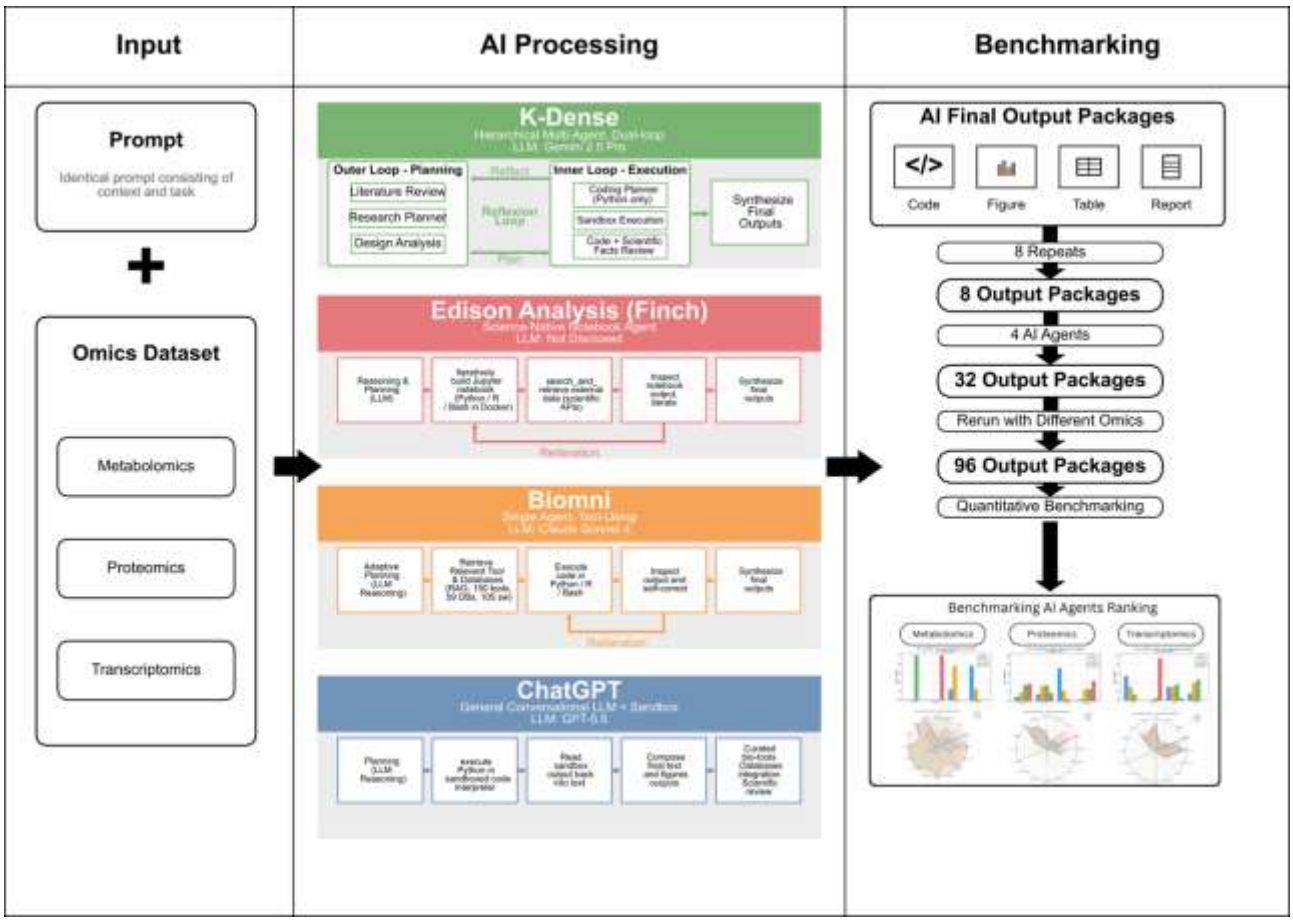


Fig. 1 | Full benchmarking framework for AI agents on multi-omics primary-data analysis. The figure depicts the four agents evaluated in this study and the rubric-based scoring stack applied to their outputs. Agent backbones and execution environments: K-Dense (built on Gemini 2.5 Pro; agentic system with iterative outer- and inner-loop reasoning), Edison/Finch (undisclosed base model; managed system), Biomni (Claude LLM-based; tool-using research assistant with code execution and database access), and ChatGPT (general-purpose LLM with a code-interpreter sandbox; no specialized bioinformatics scaffolding). Each agent received an identical prompt and was run autonomously, without human intervention, aside from a single "try again" re-prompt when an agent explicitly flagged its analysis as failed. Color code throughout the manuscript: K-Dense, green; Edison/Finch, red-pink; Biomni, orange; ChatGPT, blue. Each agent–modality replicate was run eight times (8 × 4 agents × 3 modalities = 96 runs in total across the three modalities). Statistical uncertainty was quantified using a nonparametric bootstrap (10,000 iterations; seed 42; 95% confidence intervals).

In this benchmark design, each agent received an identical prompt per omics dataset and was independently run eight times, producing a total of 32 agent-run combinations per modality and 96 across the full study (prompts provided in **Supplementary Tables S1-S3**). The prompt for each modality was deliberately designed with different instructional specificity — open-ended for metabolomics, scoped for proteomics (defined groups, comparisons, and preprocessing state), and fully prescriptive for transcriptomics (assigned tools, thresholds, exclusions, and steps). Such an experimental design addressed how the three modalities are actually queried in practice: metabolomics is dominated by exploratory biomarker discovery with weak methodological consensus, proteomics case-control studies typically present defined groups and pre-processed matrices, and transcriptomics analyses are governed by an established edgeR/DESeq2/TMM tooling set. Accordingly, the metabolomics prompt states only a research goal, an analysis scope, a methodological preference, and a success metric; the proteomics prompt fixes the comparisons, sample counts, fold-change formula, and preprocessing state; and the transcriptomics prompt mandates the full analytical cascade (filterByExpr, TMM, quasi-likelihood F-test, thresholds, tool

exclusions, bootstrap validation, and mechanistic hypothesis). Together the three prompts stress-test complementary agent capabilities, measuring both what an agent knows and how well it follows instructions. Standardizing every prompt to the same length or specificity for all three modalities would remove the variations in omics agents' application and no longer reflect how researchers actually query omics agents in day-to-day analysis.

For the metabolomics, the present study was based on a secondary analysis of de-identified metabolomic data originally generated in the study by Kuzin et al. on metabolomic predictors of atrial fibrillation in patients with coronary artery disease[15]; the dataset was provided for the present analysis by the original study team, and no additional biological samples were collected, nor additional LC–MS/MS measurements performed. Each agent received an identical prompt per omics dataset and was independently run eight times, producing a total of 32 agent–run combinations per modality and 96 across the full study (prompts provided in **Supplementary Tables S1–S3**; dataset provenance and pre-processing detailed in **Methods «Pairwise similarity, reproducibility and embeddings-space visualization**). The proteomics modality used the Johnson et al. Alzheimer's-disease brain TMT proteome (3,334 proteins × 419 dorsolateral-prefrontal-cortex samples: 230 AD, 98 asymptomatic AD, 91 controls; $\log_2$-transformed, TAMPOR-batch-corrected, regressed for age/sex/PMI) [16]; the transcriptomics arm used the TCGA-BRCA STAR-aligned integer-count matrix from the NCI Genomic Data Commons (63,856 GENCODE-v36 genes × 1,256 aliquots; ~1,109 primary tumours, ~113 normals, ~7 metastases) [17,18].

In the transcriptomics track, twelve Biomni runs were initiated, with four replicates terminated after exceeding a 90-min threshold during the bootstrap-resampling step, while the eight completed runs were retained for further benchmarking. On the same track, K-Dense's outer planning loop flagged four of eight runs as failing internal verification of the final-document generation step and required a single user re-prompt to "try again" (three runs needed one re-prompt; one needed two). These operational events were recorded as agent-level behavior rather than replicates removed from the analysis (**Methods "Omics benchmark datasets and repeated agent runs"**).

Each agent output was treated as a structured scientific document and parsed into an 87-dimensional feature representation built from format-specific text extraction, overlapping 800-token chunking with 150-token overlap, structural counts, methodological keyword dictionaries and biological keyword dictionaries (Methods "Text parsing and overlapping chunk representation - Structural, methodological and biological feature extraction"). In parallel, each text chunk was embedded into a 384-dimensional dense semantic vector using the *all-MiniLM-L6-v2* sentence-transformer[19] with hierarchical mean pooling from chunks to files, runs and agent centroids (Methods "Semantic embeddings and fallback lexical representation"). Two scoring instruments were furthermore applied to every run. First, a process-anchored composite AgentScore was defined as equation (1),

$$AgentScore_{a,r} = P_{a,r} \times (0{,}5 + \frac{0{,}30 \times D_{a,r}}{100} + \frac{0{,}50 \times R_{a,r}}{100}) \qquad (1)$$

where *P* (process quality), *R* (result reporting), and D (interpretation depth) are means of 8, 7, and 7 binary indicators, respectively, each scaled to 0-100 (Methods "Composite AgentScore"). In depth, an output that fails to document its *P* Score cannot produce a positive score despite the listing of many results, as P acts as a multiplicative gate, following the framework provided by process-anchored evaluation frameworks for biomedical AI[20]. An independent Domain Score was furthermore computed from 16 modality-specific ordinal criteria graded 0/1/2 and normalized to 0–100 (Methods "Domain-aware ordinal grading"). Within-agent reproducibility was the mean pairwise cosine similarity of run-level sentence-transformer embeddings, summarized by the reproducibility (equation (1), Methods "Pairwise similarity, reproducibility and embedding-space visualization"; **Supplementary Fig. S2-S4**).

$$Q_a = max(0, 1 - \frac{CV_a}{100}) \quad (2),$$

Ranking uncertainty was quantified by bootstrap resampling with 10,000 iterations and seed 42 (Methods “Monte-Carlo centroid uncertainty - Bootstrap ranking uncertainty and pairwise comparisons”), and ranking robustness was tested under seven alternative scoring definitions A–G (Methods "Scoring robustness, output-volume diagnostics and alternative scoring definitions").

AgentScore distributions applied a stable three-tier structure with modality-specific top ranks

Mean AgentScores sorted the four agents into the same three-tier structure across all modalities, but with modality-specific orders. We report each value as the mean AgentScore followed by the 95% bootstrap CI of the mean (**Fig. 2a**). In metabolomics, K-Dense ranked first (92.66, CI 87.34–98.12), Finch was second (90.20, CI 86.09–94.22), Biomni was third (76.47, CI 69.11–82.72), and ChatGPT was fourth (39.29, CI 25.38–50.60) (**Fig. 2a**). In proteomics, Finch ranked first (86.72, CI ≈ 82.4–91.1), K-Dense was second (83.10), Biomni was third (81.76), and ChatGPT was fourth (29.87) (**Fig. 2b**). In transcriptomics, Finch ranked first (86.72, CI 85.16–87.50), K-Dense was second (81.25, CI 76.56–85.94), Biomni was third (70.18, CI 65.13–74.20), and ChatGPT was fourth (29.33, CI 12.71–45.51) (**Fig. 2c**).

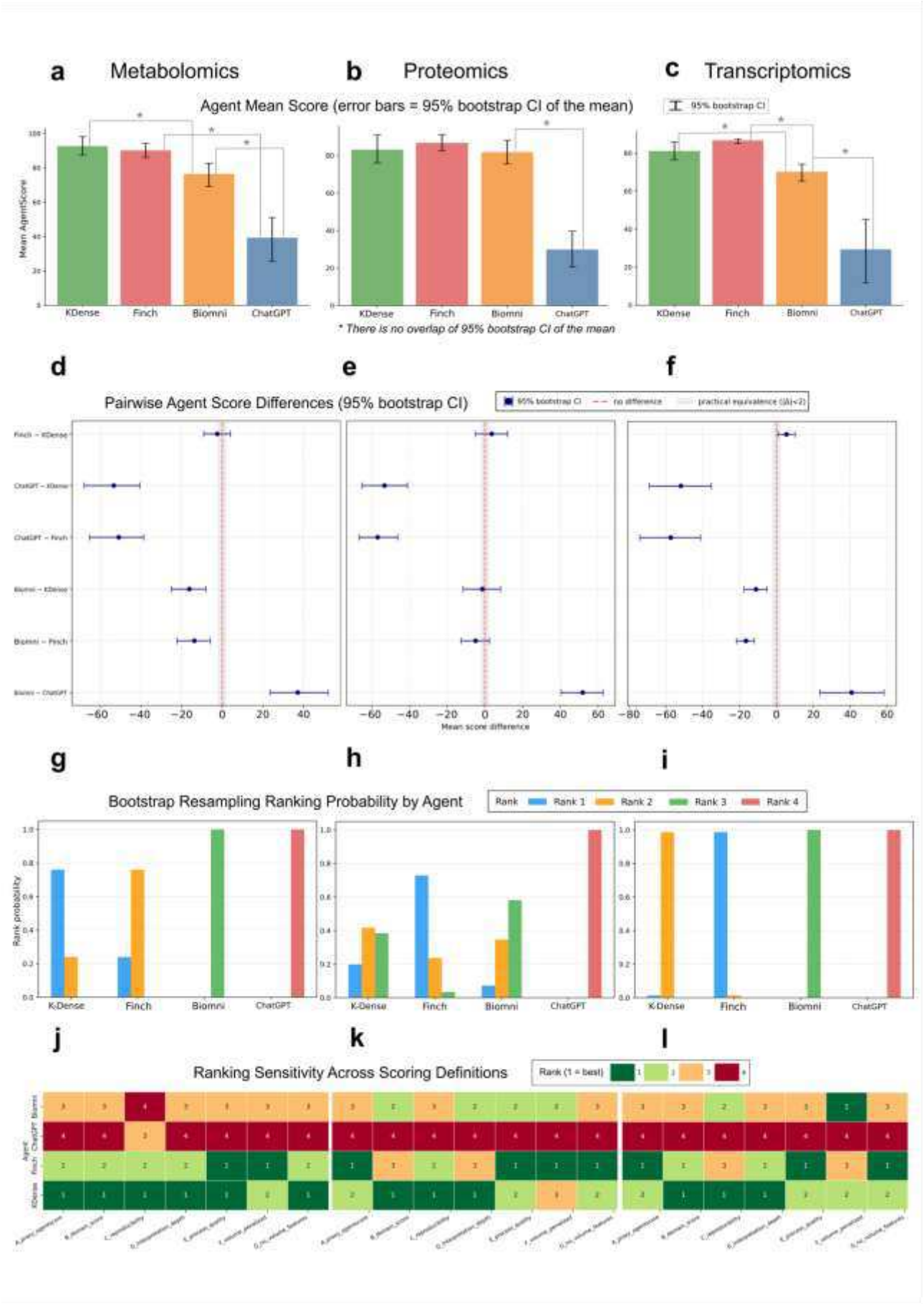


Fig. 2 | AgentScore performance across metabolomics, proteomics and transcriptomics. Quantitative comparison of the four agents on the three omics modalities. Column order: metabolomics (left), proteomics (center), transcriptomics (right). Bars and symbols are colored by

agent (K-Dense, green; Edison Analysis (Finch), red-pink; Biomni, orange; ChatGPT, blue). **a–c**, Mean AgentScore per agent. Bars, mean across n = 8 independent runs per cell; error bars, 95% bootstrap confidence intervals from 10,000 resamples (seed 42). **d–f**, pairwise mean differences between agents (forest plots). Symbols, mean difference; horizontal lines, 95% bootstrap CIs; the grey vertical band marks the pre-specified equivalence region $|\Delta| < 2$ AgentScore points (operational threshold for "indistinguishable"). **g–i**, bootstrap rank-probability bars. Each stacked bar shows the empirical probability that an agent occupies rank 1–4 across the 10,000 resamples; complete dominance corresponds to P (rank 1) → 1. **j–l**, sensitivity heatmaps of agent ranking under seven alternative scoring definitions A – G: A, primary AgentScore as in **a–c**; B, DomainScore; C, within-agent cosine similarity (reproducibility index); D, interpretation depth; E, process quality; F, volume-penalised AgentScore; G, AgentScore excluding output-volume features. Cell values denote ranks (1 = best). All statistics derive from the same n = 32 agent-run cells per modality (4 agents × 8 runs; 96 runs in total).

The findings in **Fig. 2a-c** highlighted the narrowest and widest transcriptomics CI by Finch and ChatGPT, respectively. Finch's transcriptomics CI is the narrowest in the entire study (CI half-width ≈ 1.2 AgentScore points), a magnitude tighter than the other three agents' CIs in that modality. ChatGPT's transcriptomics CI, on the other hand, is the widest in the entire study (CI half-width ≈ 16.4 points), driven by both a substantially lower mean and a larger run-to-run dispersion. Furthermore, ChatGPT's CI does not overlap any other agent's CI in any modality, isolating the bottom of the ranking from the rest with high confidence (**Fig. 2a**). The three-tier structure is therefore not an artefact of any single comparison as the top tier (K-Dense and Finch), the middle tier (Biomni), and the bottom tier (ChatGPT) are statistically separated in every modality with the only ambiguity inside the top tier itself.

The key takeaway of these distributions is that no agent is the universal winner. K-Dense leads in metabolomics, with Finch statistically tied, whereas Finch leads in proteomics, with a shared non-significant region with K-Dense and Biomni. In transcriptomics, Finch leads with a leading-pair CI that excludes zero, while Biomni occupies a stable mid-tier (mean 70–82) with no CI overlap with the top pair in any modality, but ChatGPT, in contrast, is last in every modality with non-overlapping CIs. This pattern is the primary observation of the study and motivates the architectural approach developed in section "Bootstrap rank probabilities quantify residual rank-1 ambiguity" and the multi-rubric analysis in section "Ranking sensitivity under seven scoring definitions - Domain-aware coverage, per-run component presence, and proxy x DomainScore concordance".

Pairwise mean-difference CIs

Pairwise mean-difference CIs (**Fig. 2d-f**) localize the three-tier structure of **Fig. 2a-c** and identify which agent pairs fall inside the pre-registered practical-equivalence band of $|\Delta| < 2$ AgentScore points (Methods "Bootstrap ranking uncertainty and pairwise comparisons"). A CI is judged to be inside the band only when the entire CI lies within $|\Delta| < 2$, and we report the relationship between the CI and zero separately from its relationship to $|\Delta| < 2$.

In metabolomics, Finch - K-Dense scored -2.46 with a 95% CI crossing both zero and the $|\Delta| < 2$ band, so we cannot conclude that Finch and K-Dense differ in metabolomics AgentScore (**Fig. 2d**). Every other pairwise CI excluded zero in the figure. ChatGPT - K-Dense and ChatGPT - Finch placed ChatGPT 50.91–53.3 points below the two top agents, while Biomni - K-Dense and Biomni - Finch placed Biomni 13.73–16.18 points below them, and Biomni - ChatGPT had a mean difference of +37.19 with a CI entirely outside the equivalence band.

In proteomics, the top tier comparatively compressed further. The three top-pair differences were Finch - K-Dense at +3.62, Biomni - K-Dense at -1.34, and Biomni - Finch at -4.96. All three pairs

had CIs crossing zero, placing K-Dense, Finch and Biomni in a region of mutual statistical non-significance (**Fig. 2e**). Of these three, only Biomni - K-Dense (-1.34) has a point estimate inside the $|\Delta| < 2$ band while Finch - K-Dense (+3.62) and Biomni - Finch (-4.96) has point estimates outside the band. ChatGPT remained an outlier in proteomics with ChatGPT - K-Dense at -53, ChatGPT - Finch at -56 (CIs entirely below zero) and Biomni - ChatGPT at +52 (CI entirely above zero), indicating that all three pairs are outside $|\Delta| < 2$.

In transcriptomics, the leading pair Finch - K-Dense scored +5.47 with a CI that does not cross zero (**Fig. 2f**), and the point estimate lies outside the $|\Delta| < 2$ band on the positive side. On the strict equivalence criterion, the two agents are therefore not practically equivalent in transcriptomics AgentScore even though the absolute size of the gap is small. Biomni - K-Dense scored -11 (CI ≈ [-18, -5]) and Biomni - Finch scored -17 (CI ≈ [-22, -11]), placing Biomni in the intermediate tier with non-overlapping CIs, while the two ChatGPT pairs (ChatGPT - K-Dense ≈ -52, ChatGPT - Finch ≈ -58) had CIs entirely below zero and far outside the equivalence band.

The pre-registered $|\Delta| < 2$ band is the single most consequential decision in the headline ranking, and we therefore underline the distinction between "CI crosses zero" and "CI lies inside $|\Delta| < 2$". "CI crosses zero" asks about statistical significance against the null of no difference while CI lies inside $|\Delta| < 2$" asks about practical relevance against a pre-registered effect-size threshold. Conflating them would falsely flatten a distinguishable top rank into a tie or declare practical equivalence where none has been demonstrated. We therefore report three statements separately: "CI crosses zero", "CI lies inside ±2", and "point estimate lies inside ±2". **Fig. 2d-f** makes the impact of this decision directly visible. In metabolomics, Finch - K-Dense ≈ -2.46 with a CI crossing both zero and the band, so the two leading agents are not separable under either criterion (**Fig. 2d**). In proteomics, the three top-pair differences have CIs crossing zero but only Biomni - K-Dense point estimate (-1.34) sits inside ±2 (**Fig. 2e**). The three agents are mutually non-significant against the null of no difference, but only one of the three pairs is also practically equivalent under the pre-registered band. In transcriptomics, Finch - K-Dense ≈ +5.47 has a CI that excludes zero but a point estimate outside the band (**Fig. 2f**). Thus, the leading pair is statistically distinguishable, but the absolute size of the gap is small.

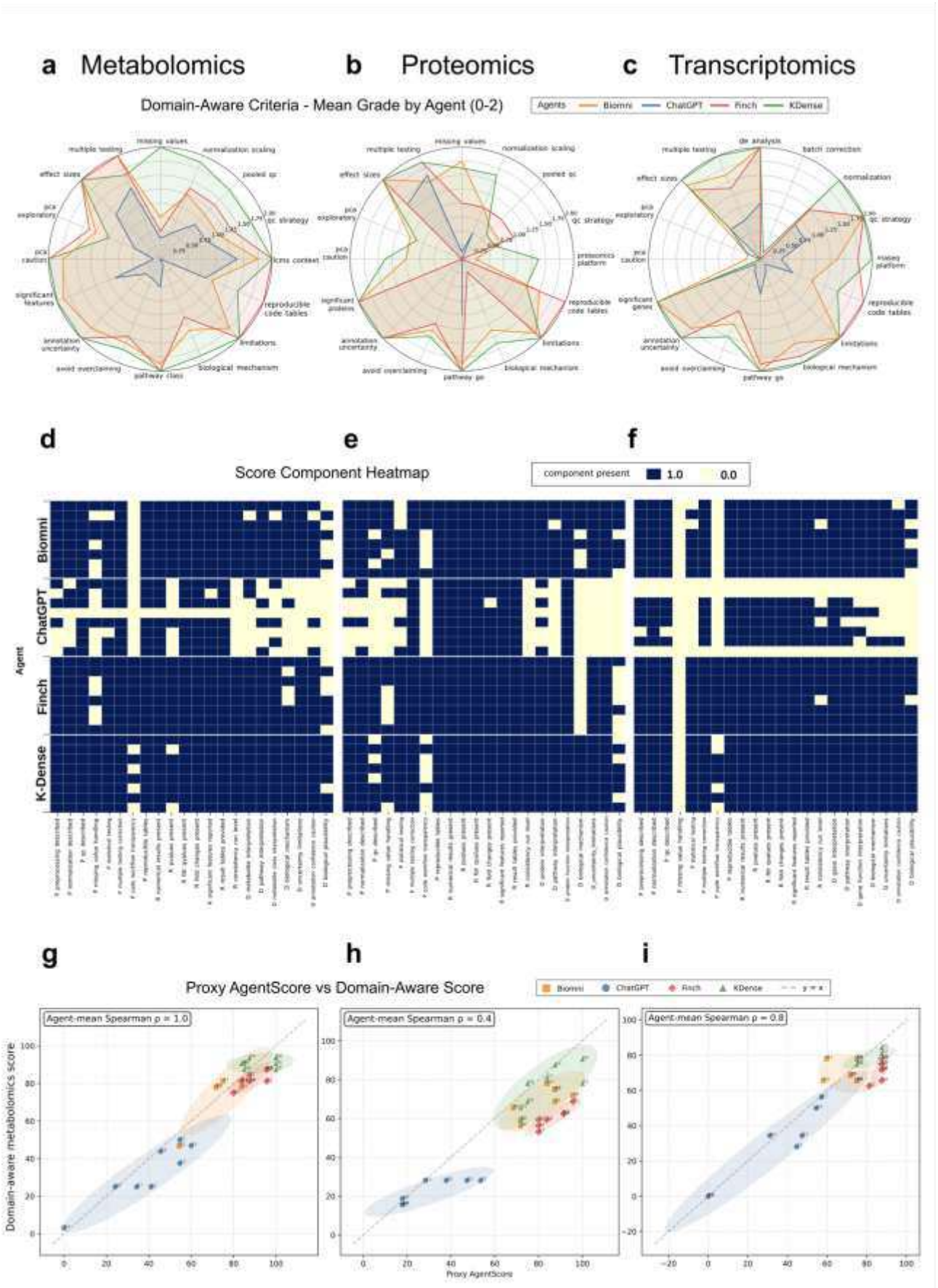


Fig. 3 | Domain-expert evaluation and within-agent reproducibility. Domain-rubric profiles, proxy-score component structure, and convergence between proxy and domain scoring. Column order matches Fig. 2: metabolomics (left), proteomics (centre), transcriptomics (right). Colour code

is identical to Figs 1–2. **a – c**, Radar profiles of the 16 modality-specific domain criteria. For each agent, the per-criterion median grade across n = 8 runs is plotted on a 0–2 scale (0, absent/incorrect; 1, partial; 2, complete and correct). **d – f**, Component heatmaps of the 22 proxy-score components (columns) across all 32 runs per modality (rows), grouped by agent. Cell colour encodes the binary or normalised component value as captured by the automated grader; row-to-row uniformity within an agent block reflects within-agent consistency. **g – i**, convergence between proxy and domain scoring. Each point is one agent-run cell (n = 32 per modality) plotted by mean AgentScore (x) versus mean DomainScore (y). Spearman rank correlation between agent-level proxy and domain scores are: $\rho = 1.0$ (metabolomics), $\rho = 0.4$ (proteomics), $\rho = 0.8$ (transcriptomics). AgentScore combines 8 process (P) binary indicators, 7 result-reporting (R) binary indicators and 7 interpretation depth (D) binary indicators; DomainScore is the mean of the 16 ordinal criteria normalised to 0 – 100. Bootstrap CIs use 10,000 resamples with seed 42.

The published statistical literature on small-sample agent evaluation has cautioned explicitly against applying central-limit asymptotics to small per-agent replicate counts [21, 22], and our reporting strictly follows that caution.

Bootstrap rank probabilities quantify

To express the rank uncertainty implicit in the overlapping CIs of section "Pairwise mean-difference CIs and the practical-equivalence band", for each agent we counted the fraction of the 10,000 bootstrap resamples in which it ended up at rank 1, 2, 3, or 4 (**Fig. 2g-i**). In metabolomics, *P* (rank 1 | K-Dense) scored 0.76 and *P* (rank 1 | Finch) scored 0.24 with Biomni at rank 3 with probability ≈ 1.0 and ChatGPT at rank 4 with probability ≈ 1.0 (**Fig. 2g**). In proteomics, the top rank was genuinely disputed, given P(rank 1 | Finch) at 0.73, P(rank 1 | K-Dense) at 0.20, and P(rank 1 | Biomni) at 0.07 with K-Dense and Biomni splitting the residual rank-2 / rank-3 mass, and ChatGPT holding rank 4 with probability ≈ 1.0 (**Fig. 2h**). In transcriptomics, the top rank was essentially fixed as P(rank 1 | Finch) scored 0.987 and *P* (rank 1 | K-Dense) scored 0.013 with Biomni at rank 3 and ChatGPT at rank 4 with probability ≈ 1.0 each (**Fig. 2i**).

Metabolomics is the only modality in which both top agents carry non-trivial rank-1 mass, consistent with the overlapping CIs in **Fig. 2d**. Meanwhile, proteomics carries the largest residual rank-1 uncertainty in the study with no agent reaching P (rank 1) > 0.73, and the same modality has K-Dense's rank-2 / rank-3 mass most evenly split. Additionally, transcriptomics shows the most concentrated rank-1 distribution in the study, consistent with Finch's unusually narrow CI in that modality (section "AgentScore distributions applied a stable three-tier structure with modality-specific top ranks"). Across all three modalities, the bottom of the ranking is fully fixed and is held by ChatGPT, a result that occurs under every alternative scoring definition we examine in section "Ranking sensitivity under seven scoring definitions".

Ranking sensitivity under seven scoring definitions

To verify the rankings of Fig. 2g–i are not an artefact of one composite score, we recomputed agent ranks under seven scoring definitions (Fig. 2j–l; Methods "Scoring robustness, output-volume diagnostics and alternative scoring definitions"). The seven definitions were selected to probe four distinct sources of potential ranking instability: the choice of scoring instrument (A vs. B), the contribution of individual score components (D, E), the effect of output volume on apparent completeness (F, G), and agent self-consistency independent of content quality (C). Specifically: A (proxy AgentScore) is the primary composite defined in equation (1) and serves as the reference; B (DomainScore) is the independent expert-graded rubric, included to test whether the automated proxy and the domain-aware instrument agree on rank order; C (within-agent reproducibility index) ranks agents by run-to-run consistency of their outputs rather than by content, isolating behavioural stability

from analytical quality; D (interpretation depth) and E (process quality) are the two multiplicative sub-components of AgentScore applied in isolation, testing whether the overall ranking is driven by one component alone; F (volume-penalised AgentScore) down-weights scores proportionally to output length, directly testing whether longer reports inflate apparent completeness; and G (AgentScore excluding output-volume features) removes all length-sensitive indicators from the feature set, providing a volume-blind complement to F. Supplementary Fig. S14–S15 show the volume diagnostics underlying definitions F and G.

The bottom of the ranking is structurally invariant in every modality. ChatGPT is ranked 4 under A, B, D, E, F and G in all three modalities with a switch probability of 0 in proteomics and transcriptomics (**Fig. 2j-l**). The single ChatGPT cell where the rank changes are reproducibility (C) in metabolomics, where ChatGPT's high cosine similarity (section "Within-agent reproducibility and its scope") places it at rank 1. This single exception is itself informative and will be discussed further in section "Within-agent reproducibility and its scope".

In metabolomics, K-Dense holds rank 1 under A, B, D, E, G and rank 2 only under F (volume-penalized), giving expected rank 1.2 and switch probability 0.24. Meanwhile, Finch's expected rank is 1.8 with the matching switch probability of 0.24 because it climbs to rank 1 under E and F. Biomni sits at rank 3 under every definition except C, where it drops to rank 4 (expected rank 3.0, switch probability 0). In proteomics, however, no agent is robustly first. Finch has expected rank 1.3 with switch probability 0.27 (falling to rank 3 under B and to rank 2 under C and D). K-Dense was expected to hold rank 2.2 with the largest switch probability we observe in the study (0.58), rising to rank 1 under B, C, D and falling to rank 3 under F. Whereas, Biomni has expected rank 2.5 with switch probability 0.42. In transcriptomics, Finch's expected rank is 1.0 with switch probability 0.01 and K-Dense's is 2.0 with switch probability 0.01, confirming the top-pair resolution seen in **Fig. 2i** while Biomni's expected rank is 3.0 with switch probability 0, the only exception being F = volume-penalised, where it rises to rank 1, consistent with Biomni's lower output volume in transcriptomics.

The headline rankings of **Fig. 2g-i** are therefore not produced by the choice of composite score. ChatGPT remains last under every alternative definition whereas the leading pair is K-Dense / Finch in metabolomics, Finch in transcriptomics, and a three-way tie of Finch / K-Dense / Biomni in proteomics, exactly the structure visible in **Fig. 2d-i**. The sensitivity analysis adds information at the top of the proteomics ranking, where the switch probabilities of 0.58, 0.42 and 0.27 quantify exactly how much of the K-Dense / Finch / Biomni order is determined by the choice of rubric rather than by agent quality. Furthermore, the proxy AgentScore and the independent domain-aware rubric disagree most strongly (section "Domain-aware coverage, per-run component presence, and proxy x DomainScore concordance"), and such observations can be explained by the agents' compression in particular AgentScore points. When the agents are compressed inside a few AgentScore points, the choice of rubric becomes the dominant determinant of the ranking, and the disagreement between rubrics is itself a structural diagnostic of the regime in which the benchmark is operating.

Within-agent reproducibility, measured as the mean pairwise cosine similarity of sentence-transformer embeddings across the eight independent runs of each agent (Methods "Pairwise similarity, reproducibility and embedding-space visualisation"; **Supplementary Fig. S2-S4**), was strongly modality-dependent (per-agent means). In metabolomics, K-Dense produced the most reproducible outputs (cosine ≈ 0.93 ± 0.02), followed by ChatGPT (≈ 0.92 ± 0.04), Finch (≈ 0.91 ± 0.03) and Biomni (≈ 0.88 ± 0.07) (**Supplementary Fig. S1**). In proteomics, K-Dense remained the most reproducible (≈ 0.94 ± 0.03), followed by Finch (≈ 0.92 ± 0.03), Biomni (≈ 0.92 ± 0.04) and ChatGPT (≈ 0.77 ± 0.13). In transcriptomics, K-Dense was again the most reproducible (≈ 0.96 ± 0.03), Biomni was second (≈ 0.87 ± 0.07), Finch was third (≈ 0.84 ± 0.10), and ChatGPT was fourth (≈ 0.53 ± 0.34).

Two patterns are worth recording from this observation. First, reproducibility and AgentScore are not interchangeable, which can be seen in ChatGPT's outputs. ChatGPT is second-most reproducible in metabolomics (cosine 0.92) but ranks last on AgentScore in the same modality (mean 39.29), emphasizing the difference of these two metrics. This combination (high reproducibility, low accuracy) is the operational signature of an agent that defaults to a small set of standard answers regardless of the data presented. Second, K-Dense is the only agent whose mean within-agent cosine exceeds 0.90 in every modality (0.93 / 0.94 / 0.96), a behavioural signature consistent with the self-reflective-agent literature on plan-and-reflect loops **(Supplementary Fig. S1)**. ChatGPT's transcriptomics cosine (0.53) is furthermore the only value in the matrix that falls below the conventional reliability threshold of 0.70 used for embedding-based agreement. Interestingly, in transcriptomics, ChatGPT is the only agent whose run-to-run stability does not meet the reproducibility criterion of the benchmark.

The two ChatGPT features are also operationally different, and we treat them differently for an accurate diagnosis. In metabolomics, ChatGPT is consistent because it defaults to a narrow set of short, standard answers across most of its runs, not because it is rigorous. In transcriptomics, by contrast, three of ChatGPT's eight runs returned zero parsable output (section "Domain-aware coverage, per-run component presence, and proxy x DomainScore concordance") and the remaining five drift in content. Its low cosine of 0.53 therefore reflects two distinct failure modes, silent failure on the failing runs and content drift on the surviving runs, rather than the consistent-shallowness signature of its metabolomics behaviour **(Supplementary Fig. S1 & S4)**. Treating these two features as equivalent because they share a metric name would lose precisely the diagnostic distinction the benchmark is designed to present.

This is also why the reproducibility-only definition (C in section "Ranking sensitivity under seven scoring definitions" places ChatGPT at rank 1 in metabolomics. The metric is doing exactly what it is designed to, which is measuring consistency. Defining a leaderboard on cosine alone (**Fig. 3j-l**) would, however, lose the diagnostic distinction between an agent that is consistent because it is shallow and one that is consistent because it is rigorous [23, 24]. Defining a leaderboard on AgentScore alone would lose the parallel signal for the operational difference between an agent that produces a stable wrong answer and one that produces unstable answers. The benchmark therefore reports both as parallel readouts and treats their disagreement as informative rather than as a contradiction to resolve.

Domain-aware coverage

We presented the rule-based domain-aware grades (0–2) as radar profiles in all three modalities (**Fig. 3a-c**; **Supplementary Fig. S6-S8**) to compare the agent outputs beyond aggregate scores. As agents differ in the analytical dimensions, agent performance is not only quantitatively but also qualitatively diverse.

The broadest and most stable radar profile was observed in the K-Dense profile across every modality. In metabolomics, it covered preprocessing, missing-value handling, significant-feature reporting, annotation uncertainty, pathway/class interpretation, biological mechanism and reproducible code/tables. In proteomics it covered platform context, normalisation, statistical reporting, significant-protein reporting, pathway/GO interpretation, limitations and reproducible code/tables. Subsequently, in transcriptomics it covered RNA-seq platform, batch correction [25], normalisation, multiple-testing correction [26], exploratory PCA, significant-gene reporting, biological mechanism, limitations and reproducible code/tables.

Finch on the other hand produced strong radar profiles with a distinctly different emphasis. Its output consistently covered limitations, annotation uncertainty, pathway-level interpretation and

reproducibility-related elements across all three modalities with weaker coverage of selected exploratory and preprocessing categories, particularly PCA-related criteria and selected upstream workflow elements in metabolomics and proteomics. Finch appeared strong in cautious scientific reporting but less consistent at reconstructing the complete analytical workflow.

Biomni interestingly showed an intermediate, more domain-dependent profile. It performed well on downstream categories: effect-size reporting, significant-feature reporting, pathway/GO or pathway/class interpretation and limitations. However, it failed to consistently cover the upstream analytical components, notably in missing-value handling in metabolomics, proteomics-specific QC elements, and batch correction as well as normalization in transcriptomics. Thus, Biomni's outputs may be insightful for biological and statistical interpretation, but further additional methodological steps are required to fully produce analysis reports.

In contrast, ChatGPT produced the narrowest radar profiles in all three modalities. Its strong coverage was concentrated in selected statistical elements (multiple-testing correction and effect-size reporting) while domain-specific preprocessing, annotation uncertainty, biological-mechanism discussion and reproducible code/tables were inconsistently represented. The benchmark indicates that its reports contained fewer of the predefined components required for a complete domain-aware omics analysis.

The observed profiling in radar plot emphasizes the need of separate benchmarking across different modalities or omics datasets. Metabolomics requires LC-MS context, missing-value handling, metabolite-annotation uncertainty, and pathway or chemical-class interpretation[27]. Proteomics demands platform awareness, protein-level reporting and pathway/GO interpretation[28]. In turn, transcriptomics demands RNA-seq-specific normalisation, batch correction, differential-expression analysis and gene-level interpretation[29]. The same agent may therefore perform differently depending on whether the task primarily demands biochemical, protein-functional, or gene-regulatory reasoning, an observation the radar profile aims to capture.

The proxy-score component heatmaps (**Fig. 3d-f**; **Supplementary Fig. S6-8**) localise these agent-level signatures at the level of individual runs and components. Each row is one of the eight independent runs per agent (32 rows = 4 agents x 8 runs, ordered top to bottom Biomni, ChatGPT, Finch, K-Dense), and each column is one of the 22 proxy-score components covering process (*P*; 8 columns), result (*R*; 7 columns) and interpretation (*D*; 7 columns). Dark blue cells mark components present while pale yellow cells mark components absent. The D-block columns are modality-specific, naming metabolites in metabolomics (*D_metabolite*), proteins in proteomics (*D_protein*) and genes in transcriptomics (*D_gene*).

K-Dense and Finch row blocks are dense and largely dark in every modality (per-block presence 89.8–96.0%), consistent with broad component coverage on a per-run basis. Biomni row blocks are also predominantly dark (86.9–90.9% per-block presence) with two characteristic gaps that recur across modalities: *P_code_workflow_transparency*, where Biomni runs are pale in metabolomics and transcriptomics (0/8 in each), and *D_biological_plausibility* in metabolomics (3/8). These are the specific components that produce Biomni's lower AgentScore relative to K-Dense and Finch.

In contrast, ChatGPT row blocks are distinctly different and reveal two noticeable structural patterns. Four of 24 ChatGPT runs (16.7%) across the three modalities returned zero of 22 components: one of eight in metabolomics (run 4) and three of eight in transcriptomics (runs 1, 2 and 8), visible in the heatmaps as fully pale rows. Additionally, no other agent has a single zero-output run across 24 runs each. Another noticeable feature, on the runs that did return outputs, ChatGPT has systematic pale-column patterns concentrated in the process-quality block and the interpretation block but with the result block largely intact in metabolomics and proteomics. In metabolomics, ChatGPT is pale in

P_preprocessing_described (3/8), *P_normalization_described* (3/8), *P_missing_value_handling* (2/8) and *P_code_workflow_transparency* (0/8); pale in most of the *D* block including *D_metabolite_interpretation* (2/8), D_biological_mechanism (1/8), *D_uncertainty_limitations* (0/8) and *D_biological_plausibility* (0/8), but remains dense in the R block (only *R_pvalues_present* at 1/8 and *R_consistency_run_level* at 0/8 are pale). In proteomics, ChatGPT is universally absent in *P_qc_described*, *P_code_workflow_transparency*, *R_consistency_run_level* and five *D* columns (*D_pathway_interpretation*, *D_biological_mechanism*, *D_uncertainty_limitations*, *D_annotation_confidence_caution*, *D_biological_plausibility*), yet matches the leading agents on the basic-statistics *R* block (*R_pvalues_present*, *R_fdr_qvalues_present*, *R_significant_features_reported* all 8/8). In transcriptomics, the same upstream-process and interpretation gaps recur (*P_missing_value_handling*, *P_statistical_testing* and *P_code_workflow_transparency* are 0/8; the *D* block sits at 1–2/8 except *D_gene_interpretation* at 5/8). The cross-modality picture on the other hand is consistent. ChatGPT's low AgentScore is not a single weak module but a systematic failure in both the process gate *P* and the interpretation amplifier *D* of the AgentScore formula. Because P multiplies the entire score, missing process components carry through with disproportionate weight; missing interpretation components further drop the *D* contribution; and the four silent-failure runs amplify both effects. *P_missing_value_handling* is universally absent across all 32 transcriptomics runs and reflects that raw RNA-seq count matrices are not typically subject to modality level rather than an agent-level signal.

The per-run heatmaps also explain the within-agent reproducibility patterns of section "Within-agent reproducibility and its scope". Among the four agents, K-Dense, Finch and Biomni each have row-to-row patterns that are visually similar within a row block. ChatGPT's row block in contrary is the most variable with three completely zero-output runs within the transcriptomics modality. Such an observation explains why ChatGPT's transcriptomics reproducibility cosine is the lowest in the study (0.53; section "Within-agent reproducibility and its scope"). Furthermore, three of eight runs are zero vectors, which depress the mean pairwise cosine.

To check whether the proxy AgentScore and the independent domain-aware rubric measure the same underlying construct, we computed Spearman ρ between proxy AgentScore and DomainScore across all run-level points within each modality (**Fig. 3g-i**). Concordance was strongest in metabolomics ($\rho = 1.0$), where every agent with a high AgentScore also had a high DomainScore and the four agent clusters occupy non-overlapping ellipses along the $y = x$ function. Concordance was substantial in transcriptomics ($\rho = 0.8$) as the four agent clusters occupy broadly the same diagonal with some within-agent dispersion at the top of the scale, particularly visible as Finch's compressed cluster around proxy ≈ 85 / domain-aware ≈ 65. Concordance was lowest in proteomics ($\rho = 0.4$). The four agent clusters shifted to overlapping, close-to-parallel bands in the upper-right quadrant with K-Dense achieving the highest DomainScores despite a proxy AgentScore comparable to Biomni and Finch.

The pattern follows the spread of the four agents on AgentScore in each modality (**Fig. 2a-c**). Concordance is highest when the agents are most separated on AgentScore (metabolomics; mean differences span 50+ points) and lowest when the agents are tightly clustered (proteomics; the top three agents are within 5 points of each other). When agents are similar in overall behaviour, the choice of scoring rubric rather than the underlying agent quality drives the ranking outcome. This empirically justifies the multi-rubric reporting strategy used throughout this study, which also explains why proteomics shows the largest switch probability in **Fig. 2j-l**. The Spearman ρ between the proxy AgentScore and the independent domain-aware rubric (**Fig. 3g-i**) thus varies from $\rho = 1.0$ in metabolomics to $\rho = 0.4$ in proteomics.

Two consequences follow. First, when deployed agents converge in headline score, no single composite is sufficient, and a benchmark must report multiple parallel rubrics. Second, the agreement between two independent rubrics is itself diagnostic of the comparison's structure. When agents are

separated by confidence intervals that exclude the pre-registered equivalence band ($|\Delta| < 2$), the two rubrics agree (high ρ) and the choice of composite is secondary — a separable regime. When agents fall inside the equivalence band, the rubrics disagree (low ρ) and the choice of composite becomes the primary determinant of the ranking — a tied regime. Transcriptomics ($\rho = 0.8$) is intermediate: the leading pair is separable while the mid-tier remains tied. We therefore recommend reporting both rubrics in every future agent benchmark, regardless of regime.

The composite AgentScore is multiplicative in *P* for a reason. The published agent-benchmark literature has repeatedly observed that an agent can produce a plausible-looking interpretation without ever documenting the analytical process [30–32]. A multiplicative gate on *P* forces an output that omits the process to be scored as such, regardless of how many interpretations it lists. The 32-component proxy feature vector and the 16-criterion domain-aware rubric were each constructed from community reporting standards for the corresponding modality [27–29] and are released as part of this benchmark (Methods "Software implementation and generated outputs"). We do not claim that the AgentScore formula is uniquely correct. Instead, we claim that publishing seven alternative scoring definitions alongside it is the minimum acceptable robustness analysis for a deployed-agent leaderboard, and that the four agents tested here remain distinguishable under all seven (**Fig. 2j-l**). Output-volume sensitivity is part of this minimum standard. Verbosity-driven bias has been documented for LLM-based preference scoring [33] and is a recognised failure mode for report-producing agents in production [24]. The two volume-controlled definitions (F = volume-penalised, G = AgentScore excluding volume features) test this directly. Rankings are essentially preserved under both. ChatGPT is ranked 4 under F and G in every modality with switch probability 0. Meanwhile, the K-Dense / Finch top-tier order in metabolomics is retained and Finch remains rank 1 in transcriptomics under F and G. The single visible sensitivity is in proteomics, where K-Dense drops to rank 3 under F (**Fig. 2k**, column F), exactly the same modality where K-Dense's switch probability is highest (0.58, section "Ranking sensitivity under seven scoring definitions"). The reasonable interpretation is that K-Dense's broad domain-aware coverage in proteomics is achieved partly through longer, more thoroughly documented reports. Reporting F and G alongside A in every future benchmark is therefore the minimum precaution against unintended verbosity incentives.

The novelty of this scoring approach lies in the process-anchored multiplicative composite combination: an independent domain-aware ordinal rubric, within-agent reproducibility as a first-class metric, bootstrap rank-probability distributions, and pairwise mean-difference CIs with a pre-registered equivalence band rather than in any individual component.

Benchmarking take away: agent harness, not base model

The benchmark evaluates the four deployed AI agents through an identical reproducible three-layers structure across each modality (**Fig. 2a-f**) with modality-specific ranking (**Fig. 2g-l**). K-Dense utilizes Gemini 2.5 Pro and the architecture comprises an inner verification loop and an outer planning loop [34]. Its behaviour aligns with the predictions of the self-reflective-agent literature as it generates the most reproducible outputs in every modality (cosine 0.93 / 0.94 / 0.96; section "Within-agent reproducibility and its scope"), the broadest and most stable radar profile (section "Domain-aware coverage, per-run component presence, and proxy x DomainScore concordance"), and the highest DomainScores in every modality (**Fig. 3a-c**). However, such an architecture presents an operational drawback as in transcriptomics, four of eight K-Dense runs required users to re-prompt after the internal verifier flagged the initial output as failing, increasing time-to-completion.

Biomni, utilizing the Claude LLM family with branching tool-use and retrieval-augmented planning applying ReAct/Toolformer agent paradigm[32, 35, 36], shows a contrasting profile. Its mean within-agent reproducibility is the lowest in metabolomics (0.88) and shows the widest cross-modality CV. Moreover, four of twelve initiated transcriptomics runs stalled at the 10,000-iteration bootstrap step

and were terminated. We do not interpret this as a model-level deficiency as the same Claude-class base model achieves high success rates in scaffolded benchmarks such as MLAgentBench [37], but as an architecture-level signature. Branching tool-use with large in-memory state has a known limiting ceiling on the heaviest analytical step (10,000-iteration bootstrap over a high-dimensional embedding matrix), consistent with the BioML-bench observation that current biomedical agents underperform on end-to-end tasks even when their underlying models are competitive [38]. In the two modalities where the bootstrap did not exceed the threshold, Biomni completed all runs and produced a stable mid-tier result across both.

Finch exhibits an architectural behaviour intersecting between K-Dense and Biomni. Its outputs are highly templated, as seen in **Fig. 2c**, where the unusually compressed transcriptomics CI of ≈ ±1.2 points reflects exactly this, and that template explains why Finch ranks first in proteomics and transcriptomics. When the domain-aware rubric weights reporting completeness heavily, an agent that produces a complete report wins. Thus, the same template also explains why Finch's per-run component heatmap shows the least row-to-row variation (**Fig. 3d-f**). Because the base model behind Finch is not disclosed, we record this as an architectural rather than a model-level inference. The present analysis cannot attribute the template signature to any particular underlying language model, and this caveat is an important consideration for any architectural comment we make about Finch.

ChatGPT, in contrast, behaves as a baseline rather than an agent. There is no domain tooling, no reiteration, and a switch probability of 0 against the bottom of every leaderboard (section 2.5). The most striking architectural consequence is that four of 24 ChatGPT runs (one in metabolomics and three in transcriptomics; section "Domain-aware coverage, per-run component presence, and proxy x DomainScore concordance") returned no parsable analysis at all, a failure mode absent in the three scaffold agents and consistent with a no-scaffold deployment that is unable to detect or recover from an off-trajectory completion.

In this benchmarking, the interaction between the scaffold around the agent's model with the analytical steps of each modality matters the most, rather than the base model. This interaction is what produces the leader-switching we observe across metabolomics, proteomics and transcriptomics, not any fixed ranking of the scaffold's architecture. Prior work on agent benchmarks has already stated that the agent harness matters more than the base model when a single task is considered in isolation[9–12, 30, 32, 39]. Our findings show that when the same four scaffolds are compared on three modalities with different analytical steps, each scaffold reveals distinctive failure mode and outputs. Thus, we point out that the interaction between scaffold and modality — not the scaffold on its own — is what decides which agent leads on any given task.

Limitations and the need for continuously updated benchmarks

The agents under this benchmarking are themselves rapidly developing subjects. ChatGPT, Biomni, K-Dense and Finch are updated on weekly-to-monthly cycles, and the version a user installs today is not the version a peer reviewer will reproduce next year. Moreover, the datasets used to benchmark agents are public, and current generations of base models have shown the capacity to memorise significant portions of public benchmarks[40, 41]. We designed specific analytical questions so that the requested comparisons and interpretations across each modality were not directly available online, but contamination remains a well-documented caveat for this study [42, 43]. Together these considerations emphasize the need for continuously updated benchmarks: scoring rubrics and prompts versioned alongside agent versions, scheduled re-runs, and public archiving of every agent run [30, 39].

Four caveats remain specific to this study. (i) Our agent set is four subjected systems; future community benchmarking should extend it. (ii) The domain-aware criteria are derived from current

community reporting standards [27–29] and will need periodic revision as those standards evolve. (iii) The reproducibility metric is text-embedding cosine, which is not sensitive to small but consequential numerical errors in tables; pairing it with run-to-run numerical concordance on a defined subset of metrics is a suggested next step. (iv) Finch's base model is not disclosed, and thus the benchmarking findings should be interpreted as harness-level rather than model-level statements. None of these caveats affects the central conclusion of this work: AI research agents perform similarly on accuracy; what really sets them apart is what they prioritise to present, so any fair evaluation has to apply multiple data types, use explicit grading rules, and report how trustworthy the scores are.

Outlook

As AI agents are increasingly integrated into biomedical and bioinformatics workflows, hybrid systems—characterized by high accuracy, a low propensity for "hallucinations," and high reproducibility—will inevitably be employed to enhance professional productivity. Our observation that no single system dominates across all omics modalities suggests that future computational pipelines will need to combine multiple sub-agents and rely on specialized architectures and tool ensembles rather than a general-purpose model.

Furthermore, methodologically, the agents included in this research differ primarily in their structural completeness and process methodology rather than simple end results; the evaluation of AI in scientific research should shift from result-oriented static leaderboards toward continuous, multi-dimensional auditing. As shown in results, the run-to-run variability proves that reproducibility cannot be assumed, suggesting that scientific workflows using AI agents will therefore require integrated replication checks and explicit uncertainty quantification as standard reporting requirements. Ultimately, our findings indicate that while AI agents can automate complex processes of analysing and interpreting large-scale omics data, their outputs should be treated as preliminary implementations that require human oversight and structured verification methods.

Methods

Study design and benchmarking objective

This study compared four available AI research agents — ChatGPT (GPT-5.5 in medium mode)[44], Biomni (Claude family)[7], Finch and K-Dense (Gemini 2.5 Pro)[45], [34] — on standardised omics analysis and interpretation tasks. Each agent received the same input data and task prompt within a given omics modality, and its outputs were collected across repeated independent runs. The resulting unit of analysis was therefore not a single answer, but an output bundle containing reports, code, tables, figures and other supplementary files. The workflow evaluated four complementary properties of each agent: methodological completeness, biological interpretation depth, reproducibility and ranking stability. We intentionally focused on process quality and analysis structure rather than claims about general intelligence or absolute biological correctness. The full computational framework is specified in the Supplementary Methods as a flow chart with labelled processing steps, mathematical definitions, input and output artefacts, and implementation notes (**Supplementary Methods**, Master sequence and Note on algorithm description and implementation; **Supplementary Fig. 1**; **Supplementary Tables 1–3**).

Omics benchmark datasets and repeated agent runs

The benchmarking framework was applied to three omics modalities: metabolomics, proteomics and transcriptomics — selected because they differ in data structure, analytical conventions, common preprocessing requirements and interpretation logic. Metabolomics requires attention to LC–MS context, missing-value handling, metabolite annotation uncertainty and pathway or chemical-class interpretation; proteomics requires platform-aware protein-level reporting and pathway or gene-ontology interpretation; transcriptomics requires RNA-seq-specific normalisation, batch correction,

differential-expression analysis and gene-level interpretation. In each modality, four agents were compared across eight independent runs, giving a designed total of 32 combinations per modality and 96 across the full study. In the transcriptomics arm, 12 Biomni runs were initiated; four were terminated as stalled after failing to make progress for more than 90 min, and the eight completed runs were retained for analysis. Biomni was run in Max mode (full tool-use and code-execution configuration) throughout all modalities. Agent-specific operational events, including terminated runs and user re-prompts, were recorded and interpreted as part of the operational behaviour of each deployed agent. A full description of the biological datasets, sample groups, agent prompts, input files and task definitions is provided in the Supplementary Methods and in the accompanying prompt files.

Output collection and file inventory

Each agent performed independent runs using a standardised prompt for the corresponding omics task. Outputs were saved in agent-specific and run-specific folders and treated as structured document bundles. The workflow supports nested folder structures such as AgentName/Run1/file.ext or AgentName/run_1/file.ext. Supported file formats included plain text, Markdown, CSV/TSV, Excel workbooks, JSON and JSONL, Jupyter notebooks, PDF, Word documents, code files, LaTeX-related files, HTML and SVG graphics. The computational pipeline first recursively scanned all output folders, inferred the agent label and run identifier from the directory structure, excluded hidden system artefacts and built a complete file inventory. For each file, the inventory recorded the filepath, filename, extension, file type, file size, agent and run identifier. Files that could not be assigned to a recognised agent or run were excluded from the main analysis. The complete ingestion procedure is described in Supplementary Methods, Section A, Algorithms 1 and 2.

Text parsing and overlapping chunk representation

Readable text was extracted from each supported file type using format-specific parsers. Plain-text and code files were read directly; notebooks were parsed by concatenating code cells and outputs; tabular files were flattened into text; Word documents were parsed paragraph by paragraph; PDFs were parsed page by page where possible; and HTML and SVG files were processed after tag removal. Files that could not be parsed were retained in the inventory but contributed zero-length text to downstream text-based analyses. Long outputs were then split into overlapping chunks to avoid treating large reports as single unstructured documents. Text was tokenised by whitespace and divided into windows of 800 tokens with 150-token overlap, corresponding to a step size of 650 tokens. This representation preserved multi-sentence analytical reasoning across chunk boundaries while standardising the scale at which semantic embeddings and feature counts were computed. The parsing and chunking procedures are specified in Supplementary Methods, Section A, Algorithms 2 and 3.

Structural, methodological and biological feature extraction

Each agent output was converted into a fixed-dimensional feature representation using structural counts, methodological keyword dictionaries, biological keyword dictionaries and audit-oriented features. Structural features included document length, numbers of tables, code blocks, P-value mentions, false-discovery-rate or q-value mentions, fold-change mentions, gene-like tokens, dataset identifiers, figures, citations, units and numeric values. Methodological features captured terms related to quality control, normalisation, scaling, batch correction, imputation, statistical testing, multiple-testing correction, effect sizes, differential expression, alignment and quantification, principal component analysis, partial least squares discriminant analysis, volcano-plot and heatmap visualisation, and pathway enrichment. Biological features captured genes, transcripts, isoforms, proteins, metabolites, pathways, gene ontology, transcription factors, immune and inflammatory

processes, cell cycle, apoptosis, metabolism, signalling, oxidative stress, biomarkers, non-coding RNA, splicing and epigenetic terms. Audit-oriented features captured preprocessing statements, significant-feature reporting, limitations, uncertainty, biological plausibility, annotation confidence, mechanistic interpretation and reproducibility language. Feature extraction was performed using case-insensitive regular-expression matching and curated keyword dictionaries. File-level feature vectors were concatenated into an 87-dimensional representation, summed to obtain run-level vectors and averaged across runs to obtain agent-level summaries. Keyword presence sets were also retained for Jaccard-based similarity analysis. The feature-extraction and aggregation procedures are described in Supplementary Methods, Section A, Algorithms 4–6.

Semantic embeddings and fallback lexical representation

In parallel with count-based feature extraction, each text chunk was embedded into a dense semantic vector using the sentence-transformer model all-MiniLM-L6-v2[19]. Chunk embeddings were L2-normalised and mean-pooled hierarchically: first from chunks to files, then from files to runs, and finally from runs to agent-level centroids. This hierarchy allowed each agent output to be compared at file, run and agent levels. If neural embeddings were unavailable, the pipeline used a lexical fallback consisting of term-frequency–inverse-document-frequency vectorisation with unigrams and bigrams followed by truncated singular-value decomposition. The fallback branch produced internally comparable dense vectors but was not used to compare results across embedding backends. The embedding generation, pooling strategy and fallback representation are specified in Supplementary Methods, Section B, Algorithms 7–11.

Pairwise similarity, reproducibility and embedding-space visualisation

Agent outputs were compared in three complementary spaces. First, cosine similarity was computed between run-level semantic embeddings. Second, Euclidean distance was computed on standardised 87-dimensional feature vectors. Third, Jaccard similarity was computed separately for methodological and biological keyword presence sets. This design allowed the benchmark to distinguish semantic similarity, structural and reporting similarity, and explicit methodological overlap. Within-agent reproducibility was defined from pairwise cosine similarities among the eight independent runs of the same agent. For each agent, the mean, standard deviation, coefficient of variation, minimum and maximum within-agent cosine similarity were computed. A reproducibility index was defined as provided in equation (2), where CV_a is the coefficient of variation of within-agent cosine similarity expressed as a percentage. Between-agent similarity was computed both from agent-centroid cosine similarity and from all cross-agent run pairs. Run-level embeddings were projected for visual inspection using principal component analysis, metric multidimensional scaling, t-distributed stochastic neighbour embedding and uniform manifold approximation and projection, where available. These projections were used descriptively to inspect clustering and outlier structure and were not used as formal hypothesis tests. The similarity, reproducibility and dimensionality-reduction procedures are specified in Supplementary Methods, Section C, Algorithms 12–17.

Composite AgentScore

Agents were ranked using a process-anchored composite score, the AgentScore. The score combines three components: process quality (P), result reporting (R) and interpretation depth (D). Each component was computed from binary indicators at run level and scaled to a 0–100 range. Process quality (P) included eight components: preprocessing described, normalisation described, quality control described, missing-value handling described, statistical testing described, multiple-testing correction described, code or workflow transparency present, and reproducible tables present. Result reporting (R) included seven components: numerical results, P values, false-discovery-rate or q-values, fold changes, significant features, result tables and run-level consistency statements.

Interpretation depth (*D*) included seven components: gene, metabolite or protein interpretation; pathway interpretation; gene-function or chemical-class or protein-class interpretation; biological-mechanism discussion; uncertainty and limitations; annotation-confidence caution; and biological plausibility. The final run-level AgentScore was defined as in equation (1)

The multiplicative structure was deliberate: an output with no documented process cannot obtain a high score solely by listing many results or interpretations. Agent-level scores were obtained by averaging run-level AgentScores across the eight retained runs. The binary component construction, sub-scores and composite AgentScore are specified in Supplementary Methods, Section D, Algorithms 18–20.

Domain-aware ordinal grading

In addition to the proxy AgentScore, we computed an independent domain-aware score for each modality. Each omics domain was represented by 16 modality-specific criteria graded on an ordinal 0/1/2 scale, where 0 indicates absent or inconsistent reporting, 1 indicates partial reporting and 2 indicates clear reporting of the expected criterion. The criteria were selected to approximate expert peer-review expectations for each modality. Transcriptomics criteria included RNA-seq platform context, quality-control strategy, normalisation, batch correction, differential-expression analysis, multiple-testing correction, effect-size reporting, exploratory principal component analysis, cautious interpretation of that analysis, significant-gene reporting, annotation uncertainty, avoidance of overclaiming, pathway or gene-ontology analysis, biological mechanism, limitations, and reproducible code and tables. Metabolomics criteria included LC–MS context, quality-control strategy, pooled quality control, normalisation and scaling, missing-value handling, multiple-testing correction, effect sizes, exploratory principal component analysis, caution in interpreting that analysis, significant-feature reporting, annotation uncertainty, avoidance of overclaiming, pathway or chemical-class interpretation, biological mechanism, limitations, and reproducible code and tables. Proteomics criteria followed the same logic but focused on proteomics platform context, protein-level reporting and pathway or gene-ontology interpretation. The 16 ordinal grades were summed and normalised to a 0–100 DomainScore. Mean grade vectors were also used to construct radar plots showing criterion-level coverage per agent and modality. The domain-aware grading system is specified in Supplementary Methods, Section D, Algorithm 21.

Statistics

Ranking uncertainty was quantified by bootstrap resampling of run-level scores. For each iteration and each agent, eight run scores were sampled with replacement and averaged; agents were then ranked within each bootstrap iteration. This produced a rank-probability distribution $\pi_a(r)$, defined as the probability that agent an occupied rank r, rather than a single deterministic leaderboard. Pairwise win probabilities were computed for every ordered pair of agents as the fraction of bootstrap iterations in which one agent's resampled mean score exceeded the others. Bootstrap standard errors and 95% confidence intervals were computed for mean scores. Pairwise score-difference distributions were summarised by mean difference, standard error and 95% confidence interval. A practical-equivalence band of $|\Delta| < 2$ AgentScore points was pre-specified for pairwise comparisons. To characterise uncertainty in agent positions in semantic embedding space, the workflow generated Monte Carlo satellite clouds around each agent centroid using two complementary procedures. First, bootstrap centroid resampling drew eight run embeddings with replacement and averaged them to generate non-parametric centroid replicates. Second, a regularised multivariate Gaussian model was fitted after principal-component pre-reduction of the run-level embedding matrix, and synthetic satellite points were sampled from the ridge-regularised covariance structure; these satellites were used as uncertainty probes around observed agent centroids and were not treated as additional real observations. Monte Carlo procedures used 10,000 iterations with random seed 42. Beyond the main

bootstrap ranking analysis, the implementation computed error summaries for bootstrap standard errors and 95% confidence intervals of score variants, AgentScore means, pairwise score differences, rank statistics, reproducibility indices, between-agent similarities and Monte Carlo centroid-distance distributions. Run-level outliers were flagged using agent-specific score distributions, with thresholds defined relative to the mean and standard deviation within each agent; these procedures were used to audit whether reported rankings were robust to run-level variability rather than driven by isolated outputs. Inter-agent reliability of the scoring instrument was assessed using Kendall's coefficient of concordance and Cronbach's α across scoring definitions and agent-level score variants. The bootstrap ranking and pairwise uncertainty framework is specified in Supplementary Methods, Section F, Algorithm 24; the centroid procedures in Section E, Algorithms 22 and 23; extended error summaries in Section J, Algorithm 27; and inter-agent reliability in Section G, Algorithm 25.

Scoring robustness, output-volume diagnostics and alternative scoring definitions

Report-producing agents can inflate apparent completeness simply by generating longer outputs: a verbose report may satisfy more keyword-based indicators not because the analysis is more rigorous, but because more text increases the probability of matching any given criterion. To test whether the rankings reported in Fig. 2g–i are robust to this failure mode and to the choice of scoring instrument more broadly, we evaluated agent performance under seven alternative scoring definitions and explicitly quantified the relationship between output volume and score.

Output volume was characterised by seven metrics — file count (number of distinct output files), text length (total character volume across all files), table count (structured tabular outputs), figure count (visual outputs), chunk count (number of 800-token overlapping windows, reflecting document granularity), total file size (storage footprint) and artefact diversity (number of distinct file formats). These seven metrics were selected because they capture verbosity along independent dimensions: a genuinely inflated output would elevate multiple metrics simultaneously, whereas a high score driven by analytical quality would not correlate uniformly across all seven. Spearman correlations between AgentScore and each volume metric were computed within each modality to test this directly.

Agent rankings were then recomputed under seven scoring definitions, selected to cover the four principal axes along which a composite leaderboard can be unstable: instrument choice, component decomposition, behavioural consistency, and verbosity. A (proxy AgentScore) is the primary composite from equation (1) and serves as the reference. B (DomainScore) is the independent expert-graded rubric; it was included because a proxy score built from keyword matching and a rubric graded by domain criteria are methodologically orthogonal — agreement between A and B confirms that both instruments capture the same underlying construct, while disagreement identifies a tied regime in which rubric choice, not agent quality, drives the ranking. C (within-agent reproducibility index) ranks agents by run-to-run output consistency rather than content; it was included because consistency and quality are conceptually distinct — an agent can be highly reproducible by always producing the same shallow answer, and C makes this failure mode visible. D (interpretation depth) and E (process quality) apply the two multiplicative sub-components of AgentScore in isolation; they were included to test whether the full composite is dominated by one component, which would mean the other adds noise rather than signal. F (volume-penalised AgentScore) and G (AgentScore excluding volume-sensitive features) were included as two mechanistically distinct verbosity controls: F tests whether penalising raw length changes the ranking, and G tests whether the same ranking shift occurs when length-dependent features are removed from the feature set entirely; divergence between F and G would indicate that verbosity bias operates through feature inflation rather than through raw output length alone. The full procedures are described in Supplementary Methods, Section I, Algorithm 26.

Consensus and divergence analysis

The workflow also performed consensus and divergence analysis to identify methodological, biological and reporting terms shared across agents and terms preferentially associated with individual agents. This analysis was used to distinguish common analytical content from agent-specific interpretative signatures. It is described in Supplementary Methods, Section K, Algorithm 28.

Software implementation and generated outputs

The workflow was implemented in Python (recommended versions 3.10–3.12). Core dependencies were NumPy, pandas, SciPy, scikit-learn, matplotlib, seaborn, plotly, tqdm, openpyxl, tabulate and markdown. Optional dependencies were sentence-transformers for semantic embeddings[19], umap-learn for uniform manifold approximation and projection, PyMuPDF and python-docx for document parsing, and Streamlit for the interactive dashboard. The pipeline generated inventory tables; file-, run- and agent-level feature tables; pairwise similarity matrices; within-agent reproducibility tables; between-agent similarity tables; AgentScore and ranking summaries; bootstrap rank-probability tables; pairwise win-probability matrices; score-error and bootstrap-error summaries; output-volume diagnostics; consensus and divergence tables; principal component analysis, multidimensional scaling, t-distributed stochastic neighbour embedding and uniform manifold approximation and projection visualisations; Monte Carlo centroid plots; radar plots; score-distribution figures; ranking-sensitivity figures; and final Markdown and HTML reports. The full output structure is described in Supplementary Methods, Section H, and the complete index of algorithms is provided in Supplementary Methods, Section L.

Data availability

The omics datasets analysed by the agents are described in detail in the Supplementary Methods, with accessions and source repositories listed therein. All AI-agent output bundles, run-level feature tables, AgentScore and DomainScore tables, bootstrap and Monte Carlo summary tables, and similarity matrices that support the findings of this study are available in the project repository (see Code availability) and as Source Data accompanying the relevant figures.

Code availability

The source code, prompts, scoring rubric, domain-aware criteria, per-run agent outputs and analysis scripts are publicly available at https://github.com/notabene-me/Comparative-benchmarking-of-AI-research-agents.git and at Zenodo[46]. The repository includes the full computational framework, the 32-component proxy rubric, the 16-criterion domain-aware rubric, the bootstrap and Monte Carlo procedures, and the dashboard for interactive inspection.

References

1. Alseekh, S., Aharoni, A., Brotman, Y., Contrepois, K., D'Auria, J. C. et al. Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. Nat. Methods 18, 747–756 (2021). https://doi.org/10.1038/s41592-021-01197-1

2. Deutsch, E. W., Bandeira, N., Perez-Riverol, Y., Sharma, V., Carver, J. J. et al. The ProteomeXchange consortium at 10 years: 2023 update. Nucleic Acids Res. 51, D1539–D1548 (2023). https://doi.org/10.1093/nar/gkac1040

3. Deutsch, E. W., Vizcaíno, J. A., Jones, A. R., Binz, P.-A., Lam, H. et al. Proteomics Standards Initiative at twenty years: current activities and future work. J. Proteome Res. 22, 287–301 (2023). https://doi.org/10.1021/acs.jproteome.2c00637

4. Gao, S., Fang, A., Huang, Y., Giunchiglia, V., Noori, A. et al. Empowering biomedical discovery with AI agents. Cell 187, 6125–6151 (2024). https://doi.org/10.1016/j.cell.2024.09.022

5. Ifargan, T., Hafner, L., Kern, M., Alcalay, O. & Kishony, R. Autonomous LLM-driven research from data to human-verifiable research papers. Preprint at https://arxiv.org/abs/2404.17605 (2024).

6. Wang, Z., Danek, B. P. & Sun, J. BioDSA-1K: benchmarking data science agents for biomedical research. Preprint at https://arxiv.org/abs/2505.16100 (2025).

7. Huang, K., Zhang, S., Wang, H., Qu, Y., Lu, Y. et al. Biomni: a general-purpose biomedical AI agent. bioRxiv https://doi.org/10.1101/2025.05.30.656746 (2025).

8. Lu, C., Lange, R. T., Foerster, J., Clune, J., Ha, D. et al. Towards end-to-end automation of AI research. Nature (2026). https://doi.org/10.1038/s41586-026-10265-5

9. Luo, E., Jia, J., Xiong, Y. et al. Benchmarking AI scientists for omics data–driven biological discovery (BAISBench). Preprint at https://arxiv.org/abs/2505.08341 (2025).

10. Mitchener, L., Laurent, J. M., Tenmann, B., Narayanan, S., Wellawatte, G. P. et al. BixBench: a comprehensive benchmark for LLM-based agents in computational biology. Preprint at https://arxiv.org/abs/2503.00096 (2025).

11. Alber, S., Chen, B., Sun, E. et al. CellVoyager: AI CompBio agent generates new insights by autonomously analyzing biological data. bioRxiv https://doi.org/10.1101/2025.06.03.657517 (2025).

12. Zhou, J., Zhang, B., Li, G., Chen, X., Li, H. et al. An AI agent for fully automated multi-omic analyses. Adv. Sci. 11, 2407094 (2024). https://doi.org/10.1002/advs.202407094

13. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X. et al. AgentBench: evaluating LLMs as agents. Preprint at https://arxiv.org/abs/2308.03688 (2024).

14. Xiao, Y., Liu, J., Zheng, Y. et al. CellAgent: an LLM-driven multi-agent framework for automated single-cell data analysis. Preprint at https://arxiv.org/abs/2407.09811 (2024).

15. Kuzin, S. V., Lozhkina, N. G., Sotnikova, M. A., Basov, N. V., Rogachev, A. D., Gaisler, E. V. et al. Potential of metabolomic analysis for searching for new predictors of atrial fibrillation in patients with coronary artery disease. Сибирский научный медицинский журнал (Siberian Scientific Medical Journal) 46(2), 97 (2026). https://doi.org/10.18699/SSMJ20260211

16. Johnson, E. C. B., Dammer, E. B., Duong, D. M., Ping, L., Zhou, M., Yin, L. et al. Large-scale proteomic analysis of Alzheimer's disease brain and cerebrospinal fluid reveals early changes in energy metabolism associated with microglia and astrocyte activation. Nat. Med. 26, 769–780 (2020). https://doi.org/10.1038/s41591-020-0815-6

17. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012). https://doi.org/10.1038/nature11412

18. Heath, A. P., Ferretti, V., Agrawal, S., An, M., Angel, C. S., Ghahramani, A. et al. The NCI Genomic Data Commons. Nat. Genet. 53, 257–262 (2021). https://doi.org/10.1038/s41588-021-00791-5

19. Reimers, N. & Gurevych, I. Sentence-BERT: sentence embeddings using Siamese BERT-networks. In Proc. 2019 Conf. Empirical Methods in Natural Language Processing and the 9th International Joint Conf. on Natural Language Processing (EMNLP-IJCNLP) 3982–3992 (Association for Computational Linguistics, 2019). https://doi.org/10.18653/v1/D19-1410

20. Kwong, J. C. C., Khondker, A., Lajkosz, K. et al. APPRAISE-AI tool for quantitative evaluation of AI studies for clinical decision support. JAMA Netw. Open 6, e2335377 (2023). https://doi.org/10.1001/jamanetworkopen.2023.35377

21. Miller, E. Adding error bars to evals: a statistical approach to language model evaluations. Preprint at https://arxiv.org/abs/2411.00640 (2024).

22. Bowyer, S., Aitchison, L. & Ivanova, D. R. Position: don't use the CLT in LLM evals with fewer than a few hundred datapoints. Preprint at https://arxiv.org/abs/2503.01747 (2025).

23. Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. Nature 630, 625–630 (2024). https://doi.org/10.1038/s41586-024-07421-0

24. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z. et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. ACM Trans. Inf. Syst. (2025). https://doi.org/10.1145/3703155

25. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom. Bioinform. 2, lqaa078 (2020). https://doi.org/10.1093/nargab/lqaa078

26. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. Ann. Stat. 29, 1165–1188 (2001). https://doi.org/10.1214/aos/1013699998

27. Kodra, D., Pousinis, P., Vorkas, P. A., Kademoglou, K. & Liapikos, T. Is current practice adhering to guidelines proposed for metabolite identification in LC-MS untargeted metabolomics? A meta-analysis of the literature. J. Proteome Res. 21, 153–162 (2022). https://doi.org/10.1021/acs.jproteome.1c00841

28. Doncheva, N. T., Schwämmle, V. & Locard-Paulet, M. Understanding data analysis steps in mass-spectrometry-based proteomics is key to transparent reporting. J. Proteome Res. (2025). https://doi.org/10.1021/acs.jproteome.5c00287

29. Kalantari-Dehaghi, M., Ghohabi-Esfahani, N. & Emadi-Baygi, M. From bulk RNA sequencing to spatial transcriptomics: a comparative review of differential gene expression analysis methods. Hum. Genomics 20, 9 (2025). https://doi.org/10.1186/s40246-025-00884-w

30. Liu, Y., Zhou, L., Du, X. et al. Benchmarking LLM-based agents for single-cell omics analysis. Genome Biol. (2026). https://doi.org/10.1186/s13059-026-03998-z

31. Chen, Z., Chen, S., Ning, Y. et al. ScienceAgentBench: toward rigorous assessment of language agents for data-driven scientific discovery. Preprint at https://arxiv.org/abs/2410.05080 (2024).

32. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I. et al. ReAct: synergizing reasoning and acting in language models. Preprint at https://arxiv.org/abs/2210.03629 (2023).

33. Saito, K., Wachi, A., Wataoka, K. & Akimoto, Y. Verbosity bias in preference labeling by large language models. Preprint at https://arxiv.org/abs/2310.10076 (2023).

34. Li, O., Agarwal, V., Zhou, S., Gopinath, A. & Kassis, T. K-Dense Analyst: towards fully automated scientific analysis. Preprint at https://arxiv.org/abs/2508.07043 (2025).

35. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M. et al. Toolformer: language models can teach themselves to use tools. Preprint at https://arxiv.org/abs/2302.04761 (2023).

36. Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K. & Yao, S. Reflexion: language agents with verbal reinforcement learning. Preprint at https://arxiv.org/abs/2303.11366 (2023).

37. Huang, Q., Vora, J., Liang, P. & Leskovec, J. MLAgentBench: evaluating language agents on machine learning experimentation. Preprint at https://arxiv.org/abs/2310.03302 (2024).

38. Miller, H. E., Greenig, M., Tenmann, B. et al. BioML-bench: evaluation of AI agents for end-to-end biomedical ML. bioRxiv https://doi.org/10.1101/2025.09.01.673319 (2025).

39. Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N. & Narayanan, A. AI agents that matter. Trans. Mach. Learn. Res. (2024). https://doi.org/10.48550/arXiv.2407.01502

40. Magar, I. & Schwartz, R. Data contamination: from memorization to exploitation. In Proc. 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 157–165 (Association for Computational Linguistics, 2022). https://doi.org/10.18653/v1/2022.acl-short.18

41. Oren, Y., Meister, N., Chatterji, N. S., Ladhak, F., Hashimoto, T. B. et al. Proving test set contamination in black-box language models. Preprint at https://arxiv.org/abs/2310.17623 (2024).

42. Xu, C., Guan, S., Greene, D. & Kechadi, M. T. Benchmark data contamination of large language models: a survey. Preprint at https://arxiv.org/abs/2406.04244 (2024).

43. Deng, C., Zhao, Y., Tang, X., Gerstein, M. & Cohan, A. Investigating data contamination in modern benchmarks for large language models. In Proc. 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) 8706–8719 (Association for Computational Linguistics, 2024). https://doi.org/10.18653/v1/2024.naacl-long.482

44. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I. et al. GPT-4 technical report. Preprint at https://arxiv.org/abs/2303.08774 (2023).

45. Gemini Team, Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A. et al. Gemini 1.5: unlocking multimodal understanding across millions of tokens of context. Preprint at https://arxiv.org/abs/2403.05530 (2024).

46. Basov, N., Tirta, Y., Li, Z. & Piatkevich, K. Omics AI agent benchmarking framework. Zenodo https://doi.org/10.5281/zenodo.20841171 (2026).

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 32571589 and W2432024 to K.D.P.); the High-Risk High-Impact Program of the School of Life Sciences, Westlake University (Grant No. WU2026B002 to K.D.P.); and the Foundation of Westlake University (to K.D.P.).

Ethics approval and consent to participate

This study was purely computational and did not involve the recruitment of human participants, interventions involving human participants or animals, or the collection of new identifiable personal data. Therefore, ethics approval, informed consent, and compliance with human-subject research standards, including the Declaration of Helsinki, were not applicable.

Consent for publication

Not applicable.

Ethics, Consent to Participate, and Consent to Publish declarations

Not applicable.

Acknowledgement
We thank Angela Yiu from Edison Scientific and Orhan Celiker from Cisco for helpful discussion. We also thank Shihao Zhou for critical feedback on the manuscript. This study is supported by the National Natural Science Foundation of China (Grant No. 32571589 and W2432024 to K.D.P.); by the High-Risk High-Impact Program (Grant No. WU2026B002 to K.D.P.) from School of Life Sciences, Westlake University; by Foundation of Westlake University (to K.D.P).

Author contributions

N.V.B., Y.K.T., Z.L. and K.D.P. conceived the study and designed the benchmarking strategy. N.V.B. developed the computational framework, implemented the benchmarking pipeline, performed the comparative analyses, interpreted the results and prepared the initial manuscript draft. N.V.B., Y.K.T. and Z.L. collected, organized and curated the AI-agent outputs used for benchmarking. Y.K.T. contributed to study design, data organization, benchmarking validation, visualization, interpretation of the results and manuscript writing. Z.L. contributed to study design, data collection, curation, validation analyses, visualization and manuscript writing. H.Z. contributed to manuscript review and editing, interpretation of the results and funding acquisition. K.D.P. supervised the study, provided overall conceptual guidance, resources and funding acquisition, contributed to study design and interpretation of the results, and critically revised the manuscript. All authors discussed the results, reviewed the manuscript and approved the final version. N.V.B. and Y.K.T. contributed equally to this work.

Competing interests
The authors declare no competing interests.

Supplementary Files

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

- SupplementaryMethods.docx
- TableS1.xlsx
- TableS2.xlsx
- TableS3.xlsx