nature medicine

https://doi.org/10.1038/s41591-026-04808-6

Accelerated Article Preview

An open agentic AI co-scientist for clinical and biomedical discovery

Received: 23 March 2026

Accepted: 10 November 2026

Accelerated Article Preview
Published online: 16 November 2026

Cite this article as: Roberts, K. F. et al. An open agentic AI co-scientist for clinical and biomedical discovery. *Nature Medicine* https://doi.org/10.1038/s41591-026-04808-6 (2026).

Open access

Kaleigh F. Roberts, Zachary B. Abrams, Luca Cappelletti, Mahdi Moqri, Nicholas Heugel, J. Harry Caufield, Mathieu Bourdenx, Yan Li, Jineta Banerjee, Luca Foschini, Diego Galeano, Nomi L. Harris, Melody Li, Kejun Ying, Justin A. Melendez, Nicolas R. Barthélemy, James G. Bollinger, Yingxin He, Vitaliy Ovod, Tammie L. S. Benzinger, Shaney Flores, Brian A. Gordon, Taylor Pedersen, Jinsheng Yu, Brendan P. Lucey, Erik S. Musiek, Sarah A. Naylor, Adegoke A. Ojewole, Mukta Phatak, Donald L. Elbert, Sarah Biber, Eric C. Landsness, Christopher J. Mungall, Randall J. Bateman & Justin T. Reese

This is a PDF file of a peer-reviewed paper that has been accepted for publication. Although unedited, the content has been subjected to preliminary formatting. Nature Medicine is providing this early version of the typeset paper as a service to our authors and readers. The text and figures will undergo copyediting and a proof review before the paper is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers apply.

ACCELERATED ARTICLE PREVIEW

Editor summary:
OpenScientist, an open-source AI co-scientist, reproduced expert findings and generated testable hypotheses across expert-evaluated biomedical studies, including air-gapped runs on open-weight models

Editor recognition statement:
Primary handling editor: Jerome Staal, in collaboration with the Nature Medicine team.

Peer review information
Nature Medicine thanks Jakob Kather, Olivier Elemento and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

1. Extended Data

Figure or Table # Please group Extended Data items by type, in sequential order. Total number of items (Figs. + Tables) must not exceed 10.	Figure/Table title One sentence only	Filename Whole original file name including extension. i.e.: Smith_ED_Fig1.jpg	Figure/Table Legend If you are citing a reference for the first time in these legends, please include all new references in the main text Methods References section, and carry on the numbering from the main References section of the paper. If your paper does not have a Methods section, include all new references at the end of the main Reference list.
Extended Data Fig. 1	Variability in proteostasis genes mentioned across 10 runs of Case Study 3.	Roberts_ED_Fig1.pdf	The heatmap shows whether each proteostasis-related gene was mentioned in a given OpenScientist run. Blue indicates mentioned; light gray indicates not mentioned.
Extended Data Fig. 2	Findings and genes investigated as a function of iteration in Case Study 3.	Roberts_ED_Fig2.pdf	Cumulative number of findings recorded in the knowledge state (KSDS) (a) and cumulative number of unique genes from the dataset named in those findings (b) after the given number of iterations (x-axis) for the ten replicate runs of Case Study 3 listed in Supplementary Table 5. Grey lines indicate individual runs. Blue lines and shaded band are the mean ± standard deviation across runs. A finding is counted from the iteration in which it was first recorded; genes are counted once, at the first finding that names them, and are restricted to gene symbols present in the dataset.
Extended Data Fig. 3	Schematic of the AI-driven hypothesis	Roberts_ED_Fig3.pdf	The study began with the OpenScientist platform analyzing the multiple myeloma dataset (n=99) to identify differentially

ACCELERATED ARTICLE PREVIEW

	generation and rigorous validation framework.		expressed genes and functional pathways across the multiple myeloma progression spectrum (NPC normal plasma cell, MGUS monoclonal gammopathy of uncertain significance, SMM smoldering multiple myeloma, MM multiple myeloma). To validate these findings, a dual-pronged approach was employed using the dataset (n=162). A negative control was established by randomizing disease labels (Dataset R) in a copy of the validation set (Dataset V). OpenScientist successfully confirmed hypotheses in the authentic data while correctly identifying the absence of reproducible signals in the randomized control, confirming the platform's analytical specificity.
Extended Data Table 1	Summary of 10 independent runs of Case Study 3 showing consistency in primary and secondary findings as well as self-raised limitations.	Roberts_ED_Table1.pdf	Benchmark run and nine replicates (Claude Opus 4.6; Methods). Run ID, OpenScientist job identifier. Main finding and Secondary, principal and supporting conclusions of each run's report (parentheses give the finding number in the knowledge state). Statistical considerations, robustness checks the agent performed unprompted. Limitations, limitations the agent stated. Blank, not addressed.
Extended Data Table 2	Scientific criteria addressed by OpenScientist per run, by agent/model configuration and network mode (online or air-gapped).	Roberts_ED_Table2.pdf	Each cell shows the number of runs (out of 10) meeting the criterion C1-C3. C1, C2 and C3 correspond to the three clauses of the research question. C1: compares AT8 (tangle-bearing) against MAP2 (tangle-free) neurons. C2: characterises how the proteostasis network is rewired, naming at least two arms. C3: addresses lysosomal acidification specifically. FDR: false discovery rate corrected for multiple testing. Donor-aware: treats the 8 donors as the sample size rather than the 22,492 cells, by pairing or pseudobulk. Neurons from one donor are not independent observations, and ignoring that treats 8 biological replicates as 22,492. The only variation is GLM 5.2's multiple-testing correction, omitted in two online runs and one air-gapped run. Run IDs lists the OpenScientist job identifiers of the ten runs in each row.

ACCELERATED ARTICLE PREVIEW

Extended Data Table 3	Common findings across human and LLM adjudication of Case Study 5 Analysis A (human) vs Analysis B (OpenScientist) with adjudicator determined overlap.	Roberts_ED_Table3.pdf	Overlap (i.e. was the finding present in both analyses) and concordance (i.e. did both analyses agree). Adjudicators varied in the granularity with which findings were enumerated and in the criteria used to match conceptually related findings across analyses, resulting in some differences in overlap ratings. Despite these differences, all three adjudicators identified concordance on the primary finding and agreed that many findings were unique to one analysis. The principal discordances between analyses reflected differences in analytical approach (e.g., pathway-level mean logFC versus individual-protein analyses) or differing interpretations of otherwise concordant data.

1. Supplementary Information:

A. PDF Files

Item	Present?	Filename Whole original file name including extension. i.e.: Smith_SI.pdf. The extension must be .pdf	A brief, numerical description of file contents. i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*
Supplementary Information	Yes	Roberts_SI.pdf	Supplementary Tables 1-10, Supplementary Figure 1, Supplementary Methods (benchmarking with open-weight models and in air-gapped mode; knowledge-state ablation; additional case studies 6-8) and Supplementary References
Reporting Summary	Yes	NMED-FT150768C RS.pdf	
Peer Review Information	No	*N/A*	

ACCELERATED ARTICLE PREVIEW

3. Source Data

Parent Figure or Table	Filename Whole original file name including extension. i.e.: *Smith_SourceData_Fig1.xls,* or *Smith_ Unmodified_Gels_Fig1.pdf*	Data description i.e.: Unprocessed western Blots and/or gels, Statistical Source Data, etc.
Source Data Fig. 2	Roberts_SourceData_Fig2.xlsx	CONSORT box counts, query-type and data-type bar-chart values, and per-run assignments for all 93 development-period runs
Source Data Fig. 3	Roberts_SourceData_Fig3.xlsx	ROC curve coordinates (sensitivity vs 1 - specificity) for the four biomarkers in the OpenScientist panel, and AUC with 95% CI for each biomarker in both the human-analyst and OpenScientist panels (panel b)
Source Data Fig. 4	Roberts_SourceData_Fig4.xlsx	Model c-index values (panel b) and published biological-clock c-index comparison (panel c)
Source Data Extended Data Fig./Table 1	Roberts_SourceData_EDFig1.xlsx	Extended Data Fig. 1: gene-by-run inclusion matrix (1 = gene named in the run findings) for the 130 proteostasis panel genes, benchmark run and nine replicates
Source Data Extended Data Fig./Table 2	Roberts_SourceData_EDFig2.xlsx	Extended Data Fig. 2: cumulative findings and cumulative dataset genes per iteration for the ten replicate runs, with per-iteration mean and SD

ACCELERATED ARTICLE PREVIEW

An open agentic AI co-scientist for clinical and biomedical discovery

Kaleigh F. Roberts[1,2,3]*, Zachary B. Abrams[3,4]*, Luca Cappelletti[5], Mahdi Moqri[6], Nicholas Heugel[2,3,7], J. Harry Caufield[8], Mathieu Bourdenx[3,9], Yan Li[7], Jineta Banerjee[10], Luca Foschini[10], Diego Galeano[11,12,13], Nomi L. Harris[8], Melody Li[2,7], Kejun Ying[14,15], Justin A. Melendez[2,3,7], Nicolas R. Barthélemy[2,7], James G. Bollinger[2,7], Yingxin He[2,7], Vitaliy Ovod[2,7], Tammie L. S. Benzinger[16], Shaney Flores[16], Brian A. Gordon[16], Taylor Pedersen[17], Jinsheng Yu[18], Brendan P. Lucey[19], Erik S. Musiek[7], Sarah A. Naylor[2,3,7], Adegoke A. Ojewole[3], Mukta Phatak[3,20], Donald L. Elbert[3,21], Sarah Biber[3,4,7], Eric C. Landsness[3,7], Christopher J. Mungall[8], Randall J. Bateman[2,3,7]#, Justin T. Reese[3,8]#

1 Department of Pathology & Immunology, Washington University in St. Louis, St. Louis, MO, USA

2 Tracy Family SILQ Center, Washington University in St. Louis, MO, USA

3 Consortium for Biomedical Research & AI in Neurodegeneration (C-BRAIN)

4 Institute for Informatics, Data Science, and Biostatistics, Washington University in St. Louis, St. Louis, MO, USA

5 Department of Biology, University of Fribourg, Fribourg, Switzerland

6 Department of Medicine, Harvard Medical School, Boston, MA, USA

7 Department of Neurology, Washington University in St. Louis, St. Louis, MO, USA

8 Division of Environmental Genomics and Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

9 UK Dementia Research Institute at University College London, London, UK

10 Sage Bionetworks, Seattle, WA, USA

11 Electronics and Mechatronics Department, Facultad de Ingenieria, Universidad Nacional de Asuncion, Paraguay

12 Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA

13 Tesabio.ai, Boston, MA, USA

14 Department of Neurology and Neurological Sciences, Stanford University, Stanford, CA, USA

15 Institute for Protein Design, University of Washington, Seattle, WA, USA

16 Department of Radiology, Washington University in St. Louis, St. Louis, MO, USA

17 Department of Psychological and Brain Sciences, Washington University in St. Louis, MO, USA

18 Department of Genetics and McDonnell Genome Institute, Washington University in St. Louis, MO, USA

19 Department of Neurology, University of Minnesota Medical School, Minneapolis, MN, USA

20 The 10,000 Brains Project, Oakland, CA, USA

21 Department of Neurology, University of Washington, Seattle, WA, USA

* equal contribution, # corresponding authors (correspondence to batemanr@wustl.edu, justinreese@lbl.gov)

Abstract

Biomedical research is constrained by the time and domain expertise required to analyze increasingly large and complex datasets. Here we present OpenScientist, an open-source agentic AI co-scientist designed to address this gap by investigating scientist-defined queries and generating verifiable, data-grounded analyses. OpenScientist was evaluated by domain experts across five biomedical case studies, completing analyses in minutes that otherwise would take months of human time and expertise. It accurately quantified Alzheimer's blood biomarker performance, generated a plasma proteomic survival model that predicted survival in a

ACCELERATED ARTICLE PREVIEW

hidden-evaluation cohort, proposed a reproducible mechanism linking tau pathology to altered lysosomal acidification, generated multiple myeloma hypotheses validated in an external cohort with attenuated signal in randomized controls, and demonstrated concordance with an independent human analysis in a blinded, prespecified proteomic comparison. These results support the use of OpenScientist to assist biomedical data analysis in concert with human experts. To support research involving sensitive clinical datasets, OpenScientist can operate in a fully air-gapped environment using open-source models for inference. OpenScientist demonstrates that open, auditable, agentic AI can support real-world clinical research by researching the literature, generating hypotheses, executing analyses, and generating insights from complex datasets.

Introduction

The accelerating pace of biomedical data generation has created an urgent need and opportunity for computational tools that can help extract mechanistic hypotheses from complex datasets. Recent advances in large language models (LLMs) and agentic methods have enabled the development of AI-powered research assistants capable of generating hypotheses, designing experiments, and interpreting results. However, most AI scientist platforms remain proprietary, limiting reproducibility, validation, and adoption.

AI models have demonstrated impressive capabilities across scientific domains, exemplified by GraphCast for weather prediction and AlphaFold for protein structure.[1,2] More recently, AI scientific research platforms including Google's Co-Scientist[3], Sakana's AI Scientist[4], OpenAI's GPT-Rosalind[5], Edison Scientific's Kosmos[6,7], Biomni[8], K-Dense[9], and Anthropic's Claude Science[10] have emerged to support hypothesis generation, multimodal data interpretation, and scientific reasoning.[11–13] Complementary frameworks such as Paper2Agent transform published articles into scientific agents, expanding the tools for AI-assisted discovery.[14] In parallel, agentic coding systems such as Claude Code and Codex can write, debug, and execute scientific analysis code autonomously. Despite these advances, existing platforms are predominantly closed-source, preventing independent verification, customization for domain-specific workflows, and integration with institutional infrastructure.

Here, we present OpenScientist, an open-source platform to democratize AI-assisted scientific research (Fig. 1). Unlike proprietary systems, OpenScientist provides complete transparency in its workflow, enabling researchers to understand, validate, and modify the AI's reasoning. The platform is largely agnostic to data types, accepting diverse data including electronic health records, genomics, transcriptomics, proteomics, metabolomics, structural biology, histopathology, imaging, and biomarker data, in multiple input formats including text, tabular data, binary omics formats, and images. OpenScientist formulates testable hypotheses, executes computational analyses, extracts findings, and proposes mechanistic explanations grounded in data and published literature.

OpenScientist leverages Agent Skills and the Model Context Protocol (MCP) to provide domain-specific tools for data analysis, literature search, and knowledge synthesis. The system operates through iterative analysis: in each iteration, OpenScientist generates hypotheses based on accumulated knowledge, designs and executes computational experiments using sandboxed Python code execution, queries PubMed, and updates its knowledge state with validated findings. This approach mirrors the scientific method while maintaining auditability through state tracking and transparent logging.

Its modular skills system separates domain-agnostic workflow skills (hypothesis generation, result interpretation, etc.) from domain-specific analytical skills (metabolomics, genomics/transcriptomics, structural biology, statistics, etc.) allowing researchers to extend it for specialized applications without modifying core system logic.

ACCELERATED ARTICLE PREVIEW

We demonstrate OpenScientist's capabilities across clinical and biomedical domains and discuss how its open-source nature addresses needs for transparency, auditability, security, reproducibility, and customization in AI-assisted research.

Results

We evaluated OpenScientist across several biomedical use cases to assess its ability to analyze data, identify and incorporate relevant literature, interpret scientific findings, and synthesize conclusions. Five featured case studies demonstrate 1) carrying out a specified analysis plan, 2) supervised survival modeling, 3) hypothesis investigation, 4) hypothesis generation and validation in an external dataset with a negative control, and 5) prespecified analysis with expert human-vs-AI adjudication (Table 1). Three further use cases spanning 6) rare-disease transcriptomics and target discovery, 7) placebo-versus-drug adverse-event frequency analysis in clinical trial data, and 8) proteomics-based machine-learning interpretation for CSF amyloidosis prediction are reported in the Supplementary Information (section "Additional case studies"). For transparency, a summary of all queries run in OpenScientist during the testing period and the criteria used to select featured case studies are outlined in Fig. 2 and Supp. Table 1.

Case Study 1: SEABIRD Alzheimer's disease blood biomarkers in a community cohort

The first clinical study analysis consisted of multimodal clinical, plasma biomarker, and imaging data from 325 participants in the Study to Evaluate Amyloid in Blood and Imaging Related to Dementia (SEABIRD, NCT03899844) who underwent blood collection and amyloid PET imaging (Methods).[15] This clinically realistic, well-phenotyped cohort is suitable for benchmarking OpenScientist's ability to integrate heterogeneous biomedical data and investigate relationships between blood biomarkers, cognition, genetic risk, and imaging-defined pathology.

OpenScientist was asked to conduct a full analysis of the data based on the study's grant application (Fig. 3a; Methods). OpenScientist first performed exploratory data analysis to characterize cohort composition, missing data patterns, and the distributions of plasma biomarkers, demographic variables, and cognitive measures. OpenScientist reconstructed the intended analytical pipeline from the grant application, implementing receiver operating characteristic (ROC) analyses to evaluate the discriminative performance of individual plasma biomarkers for amyloid PET status, generating executable Python code for data preprocessing, feature selection, and ROC curve estimation. Verification of each generated script and output demonstrated 100% concordance between the code and reported results, although this assessed computational reproducibility rather than the validity of upstream analytical choices.

Consistent with non-blinded human-led analyses of the SEABIRD cohort, OpenScientist identified plasma %pTau217 as the highest-performing biomarker for distinguishing amyloid-positive from amyloid-negative participants (Fig. 3b).[16] In both analyses, %pTau217 substantially outperformed %pTau181, %pTau205, and Aβ42:40, with ROC curves demonstrating superior sensitivity at low false-positive rates. AUC values were identical to two decimal places for all biomarkers except pT205/T205, which had an AUC of 0.68 in the human analysis and 0.67 in the OpenScientist analysis. Both analyses used identical participant data, biomarker measurements, and outcomes, differing only in software implementations (SAS versus Python).

During this exploration phase, we executed the SEABIRD case study in OpenScientist ten times while refining our use of the interface. Across these runs, we identified several discrepancies in data handling, statistical reporting, and cohort definition. Each was traceable to ambiguity in prompt specification or assumptions made during data ingestion and preprocessing. For example, OpenScientist initially treated empty cells in CSV files as zero values rather than missing data, misclassifying participants with absent amyloid PET data as amyloid negative. In other instances, the system failed to consistently report sample sizes, confidence intervals, or

ACCELERATED ARTICLE PREVIEW

statistical methods, used a literature-derived Centiloid cutoff rather than the prespecified study threshold, and generalized findings to the full SEABIRD cohort despite the dataset representing only the confirmatory PET subset. Duplicate records in the initial data upload were erroneously treated as independent observations. These errors were not merely discrepancies in reporting and if undetected could have altered the scientific conclusions. These issues were resolved through iterative prompt refinement and data curation (Methods). Fig. 3 shows the final optimized run; a table summarizing the discrepancies across runs and the corrective modifications is provided in Supp. Table 2.

This iterative process highlights a lesson of agentic AI deployment in biomedical research: reliability depends not only on the model's analytic capacity, but also on human expertise to validate these systems with explicit instructions, transparent preprocessing logic, and structured validation. Several errors in this case study were identified because investigators had detailed knowledge of the study design and analytical framework, a level of ground truth that may not be available during novel discovery. In that setting, validation cannot depend on an expected result; users must audit intermediate steps, including cohort construction, missing-data handling, variable definitions, preprocessing decisions, statistical assumptions, and executable analysis code. Internal consistency checks, sensitivity analyses, and independent replication provide additional safeguards when the correct result is unknown. Thus, OpenScientist outputs should be treated as auditable scientific analyses requiring verification rather than as intrinsically reliable conclusions.

Case Study 2: Survival prediction by human plasma proteomics

The second evaluation involved plasma proteomic and demographic data from 500 participants spanning ages 18 to 99 years generated for the Biomarkers of Aging benchmarking challenge (Methods).[17] This well-characterized, moderately sized cohort is suitable for benchmarking survival modeling approaches and assessing the added value of high-dimensional proteomic features beyond basic demographics. OpenScientist was prompted to build a survival prediction model maximizing concordance index (c-index) on a held-out evaluation set (Fig. 4a). The c-index is a standard measure of how well a model ranks individuals by risk, where 0.5 corresponds to chance and 1.0 to perfect ordering of survival times. This design mirrors real-world applied data science workflows and aligns with the official evaluation criteria of the benchmarking challenge.

OpenScientist's primary finding was a substantial, consistent improvement in survival prediction when proteomic features were incorporated (Fig. 4b). The full proteomic model achieved a c-index of 0.796 on the internal 50:50 test split, compared to 0.615 for the age and sex baseline: an absolute improvement of 0.181 and a relative improvement of ~30%. This gain is large for survival modeling and indicates meaningful additional prognostic information in the proteomic data. Five-fold cross-validation yielded comparable c-indices of 0.811 for the proteomic model and 0.610 for the baseline, with a statistically significant paired comparison. Bootstrap resampling across 100 random splits showed the proteomic model outperforming the baseline in all iterations, with a mean improvement of 0.200 and a narrow confidence interval. These analyses reduce concerns about overfitting or favorable data partitioning. In a separate strict rerun (Methods), the frozen six-protein molecular score yielded a c-index of 0.826 on a hidden evaluation subset of 176 participants with 17 deaths (Supp. Table 3).

OpenScientist also derived a reduced nine-marker risk score that retained strong predictive performance: a c-index of 0.740 and an 8.6-fold difference in risk between lowest and highest risk groups. This parsimonious model highlights potential translational value and demonstrates that much of the predictive signal can be captured with a limited subset of markers.

OpenScientist correctly merged the two proteomic panels with metadata and identified predictors and survival outcomes. Paired statistical tests to compare model performance were methodologically sound, given that models were evaluated on identical data splits. Visualization outputs, including survival curves and risk score relationships with age, were informative and aligned with best practices, although ROC or time-dependent

ACCELERATED ARTICLE PREVIEW

AUC plots were absent. Some technical issues occurred in early iterations, including timeouts during elastic net fitting and a variable scoping error when fitting the baseline Cox model. These did not materially affect the final conclusions but highlight practical constraints rather than conceptual flaws in the approach.

OpenScientist identified biologically coherent mortality predictors clustering into distinct pathways: inflammaging (IL-6, IL-33, SPP1/osteopontin, CALCA), neurodegeneration (NEFL, NGF), organ damage (HAVCR1/KIM-1), and immune dysregulation (CD274/PD-L1). Protective factors included markers of T-cell function (CD3E, GZMA) and neural health (CNTN2, ACHE). Key predictors such as NEFL, SPP1, and HAVCR1 were independently supported by prior literature linking them to aging, systemic disease burden, and mortality. Notably, OpenScientist surfaced one highly relevant publication on plasma NEFL and all-cause mortality that had not been identified in the manual review.[18] However, several key publications linking top-ranked markers to longevity were missed, including recent work implicating CX3CL1 in human longevity, which was not indexed in PubMed at the time of analysis.[19] This highlights a dependency on literature source coverage rather than analytical reasoning and underscores the importance of complementary literature retrieval strategies when interpreting AI-generated mechanistic narratives.

Some mechanistic interpretations extended beyond the current consensus in aging biology. Directional links between immune checkpoint signaling and core aging hallmarks were presented with more confidence than supported by the literature. Additionally, the importance of certain predictors, such as CNTN2, was overstated compared to their ranking in the primary statistical models. These issues reflect challenges in balancing hypothesis generation with evidentiary caution and emphasize the need for human oversight in interpreting mechanistic narratives.

Overall, OpenScientist demonstrated strong performance across evaluation criteria. The analyses were largely correct, comprehensive, verifiable, and explainable, and reached conclusions consistent with expert analyses while providing additional robustness metrics and biological insights. According to expert review, “OpenScientist’s model performance was comparable with the top models submitted by participants in the benchmarking challenge using the same data.”[17] Fig. 4c displays the c-index from OpenScientist’s internal analysis alongside published Biomarkers of Aging clocks evaluated in the same cohort.[20–33] OpenScientist replicated and extended a complex survival modeling workflow in substantially less time than a human-led analysis, at minimal marginal cost. The primary value lies in rapid model development, systematic validation, and literature synthesis, with manageable risks related to overinterpretation rather than statistical error.

Case Study 3: Neurofibrillary tangle transcriptomics in human brain to investigate lysosomal acidification

The third evaluation involved publicly available single-soma transcriptomic data from post-mortem human brain tissue (Methods).[34] This dataset enables direct comparison of transcriptional programs between tangle-bearing neurons and matched tangle-free neurons from the same donors, providing a unique view of cellular responses to tau pathology. OpenScientist was prompted to investigate rewiring of the proteostasis network with the appearance of tau pathology, focusing on lysosomal acidification. The goal was to test OpenScientist’s ability to build a mechanistic model from an open-ended biological question. This topic was not systematically examined in the original publication, except for a mention of "autophagy" (one component of the proteostasis network), so the query required OpenScientist to interrogate the dataset in a novel direction rather than reproduce prior findings. Our prompt specified a single biological system (proteostasis) and a specific subprocess of interest (lysosomal acidification), but left the analytical approach, pathway and gene selection, and interpretation to the agent (prompt refined over three exploratory runs; Supp. Table 4).

After data loading and exploration, OpenScientist assembled a list of genes covering the arms of the proteostasis network (autophagy, lysosomal, vacuolar-ATPase - vATPase subunits, chaperones, proteasome subunits, and ubiquitin and related genes).[35,36] OpenScientist identified that several subunits of the lysosomal

ACCELERATED ARTICLE PREVIEW

vATPase (responsible for acidification) were significantly upregulated, leading to literature searches about the role of vATPase subunits in lysosomal function and neurodegeneration and links with tau clearance and aggregation. Through several iterations, OpenScientist generated the hypothesis that altered lysosomal acidification associated with tau pathology may reflect an imbalance between H^+ influx and efflux, given the increased vATPase expression and downregulation of channels such as TPC1/2, MCOLN1-3, or TMEM175 (Fig. 5).

The main findings were independently verifiable (r=0.983 correlation between claimed and verified log2 fold changes computed by a domain expert). The statistical approach was standard but did not account for the hierarchical structure of the data (8 donors, multiple cells per donor). OpenScientist noted this limitation and assessed cross-donor consistency, but did not suggest more sophisticated approaches (e.g. mixed-effect models). In subsequent analyses using newer models (Opus 4.6), OpenScientist implemented more rigorous approaches such as pseudobulking, or mixed-effect models, and noted the cross-sectional design as a possible confounding factor (Extended Data Table 1). The proposed hypothesis is consistent with current knowledge about mechanisms controlling lysosomal pH, in particular the imbalance between H^+ influx via the vATPase and H^+ leak via cation channels such as TMEM175 or potentially MCOLN1 (TRPML1).[37]

Across ten repeated runs (Methods), all analyses examined lysosomal acidification, detected V-ATPase upregulation, identified the same top hit, and inferred impaired lysosomal acidification (Extended Data Fig. 1). Interestingly, we observed that the breadth of findings and the number of genes investigated increased considerably after repeated iterations within a run, and some criteria and important findings were only addressed after several iterations (Supp. Table 5, Extended Data Fig. 2). We further observed that while the KSDS provided a succinct catalog of previous iterations which the agent uses for proceeding with subsequent analyses, ablating the KSDS did not have a deleterious effect on performance; the logs and other artifacts from which the knowledge state is derived can provide sufficient information for the analysis to proceed normally (Supp. Table 6).

The central findings were also reproducible using open-weight models Kimi K3 and GLM 5.2 and secure air-gapped mode (total: 60 runs; Extended Data Table 2; Methods). Every run identified broad induction of the proteostasis network, the paired donor structure, and the need for direct functional measurements of lysosomal pH. Despite this consistency, the genes selected or emphasized varied across runs and configurations, and the proposed mechanisms differed.

This analysis had limitations. First, despite an initial large-scale investigation, OpenScientist mostly focused on lysosomal acidification (only part of the prompt) and autophagy without investigating other aspects of the proteostasis network. This could be overcome by refining the prompt or adopting the scientist-in-the-loop approach allowing feedback during the analysis, since added as a feature (co-investigate mode). Additionally, some pathway interpretations relied heavily on a small number of significant genes, an overly reductionist approach. Lastly, OpenScientist tended to over-rely on statistical significance over effect size (e.g. TMEM175 was included in the final model despite a Log2 fold change of -0.018 because it reached statistical significance; Supp. Figure 1).

OpenScientist thus demonstrated strong performance across all criteria and reached conclusions comparable to several days of expert work within 10 iterations. Despite the limitations above, OpenScientist proposed a relevant mechanism aligning with current literature and going beyond the initial prompt. This case highlights the relevance of AI co-scientists for generating new hypotheses from latent data (published, well-curated datasets harboring analytical potential exceeding the scope of their original investigations).

Case Study 4: Multiple myeloma transcriptomics for hypothesis generation and validation

Phase 1: Exploratory analysis and hypothesis generation

ACCELERATED ARTICLE PREVIEW

The fourth case study used multiple myeloma (MM) gene expression data and tested OpenScientist's ability to: (1) generate novel hypotheses from microarray data (n=99 samples, 33,297 genes) spanning myeloma progression from normal plasma cells through monoclonal gammopathy of uncertain significance (MGUS) and smoldering MM (SMM) to active MM; (2) independently validate these hypotheses with an external cohort (Methods); and (3) identify when validation fails using randomized negative controls (Extended Data Fig. 3a). This tests whether autonomous AI systems can perform scientific validation without human supervision.

OpenScientist's initial analysis of the López-Corral et al. (2014) dataset identified 19,438 differentially expressed genes (58.4% of genome, FDR<0.05) organized into four transcriptional programs.[38] OpenScientist proposed a novel unfolded protein response (UPR) failure model, positing that MM progression involves three phases: compensated endoplasmic reticulum (ER) stress at MGUS (218 genes showing +16.3% upregulation), catastrophic UPR collapse at the MGUS to SMM transition (81.4% of gene changes), and proteotoxic crisis in active MM. Critically, OpenScientist identified a regulatory balance shift where the ratio of upregulated to downregulated genes crosses parity between SMM and MM, suggesting a "point of no return" in malignant transformation.

Phase 2: Experimental validation design

We validated the hypotheses in an external cohort (Dataset V) and in a copy of it with disease stage labels randomized as a negative control (Dataset R; Methods), testing whether OpenScientist could discriminate true biological validation (Dataset V) from spurious patterns in randomized data (Dataset R), a fundamental requirement for autonomous scientific reasoning.

OpenScientist was provided with four hypotheses derived from its López-Corral analysis: (H1) the MGUS to SMM transition represents the critical "point of no return" in disease progression, (H2) UPR failure mechanisms drive malignant transformation, (H3) widespread differential gene expression characterizes disease transitions, and (H4) specific drug targets can be identified from expression profiles.

Phase 3: AI validation results

OpenScientist correctly rejected H1 in both datasets: the Normal to MGUS transition showed substantially larger transcriptional reprogramming (615 DEGs in Dataset V, 112 DEGs in Dataset R) compared to the MGUS to SMM transition (268 and 27 DEGs, respectively). This occurred despite H1 being OpenScientist's original prediction, indicating it prioritizes evidence over self-consistency, a critical attribute for scientific discovery. The rejection was appropriately replicated across both datasets, with Normal to MGUS consistently ranking as the largest transcriptional change regardless of technical platform or data quality. This result suggests that the initial transformation to the precursor MGUS state, rather than progression to overt malignancy, may constitute the major molecular inflection point in MM pathogenesis. Willingness to reject a hypothesis when confronted with contradictory evidence demonstrates genuine hypothesis testing rather than confirmation bias.

OpenScientist's validation outcomes diverged appropriately between the two datasets, providing evidence for signal discrimination. In Dataset V, three hypotheses were supported: H2 (UPR failure mechanism with XBP1 ρ=-0.216, p=0.006; BiP ρ=-0.253, p=0.001), H3 (6,653 total DEGs across all pairwise comparisons), and H4 (BCMA $\log_2$FC=+0.92, p=0.0006; PSMB5 $\log_2$FC=+0.84, p<0.0001). In contrast, Dataset R showed no support for these hypotheses: UPR genes showed no significant changes (all p>0.2), only 582 total DEGs were identified, and differential expression analysis was not statistically significant for key drug targets. This differential validation performance demonstrates appropriate statistical stringency. The platform did not simply validate hypotheses indiscriminately; rather, it applied consistent statistical criteria (FDR correction, effect size thresholds) that rejected spurious patterns in the randomized dataset while detecting genuine signals in the true data. The contrast is particularly striking for H3 (widespread differential expression): 100% of pairwise

ACCELERATED ARTICLE PREVIEW

comparisons in Dataset V showed >50 DEGs, compared to only 40% in Dataset R, a quantitative discrimination of signal quality.

Because label randomization removes samples from their stage assignments, every DEG it yields is a false positive by construction. This reframes the H3 contrast: treating ">50 DEGs" as a hypothesis-level decision rule makes 40% a type I error rate high enough to support H3 in signal-free data. At the gene level, the 582 DEGs from Dataset R are the pipeline's noise floor; as a permutation FDR estimate (582/6,653), ~8.7% of Dataset V's DEGs may reflect noise rather than biology. Permutation analysis (Methods) gave an FDR of 7.5% for DEGs, 34.5% for stage comparisons, and 15.9% for evaluations of H3 using this >50 DEG criterion. OpenScientist computed both Benjamini-Hochberg and Bonferroni multiple test corrections in iteration 1, obtaining 0 DEGs in Dataset R, but neglected to use this analysis in the final report. This failure mode is thus both a failure to apply statistical rigor and a failure in the generation of the final report to include relevant analyses from the run.

OpenScientist identified problems with Dataset R without knowledge of the randomization. It reported 6.9-fold lower signal-to-noise ratio compared to Dataset V and flagged data processing that may have "obscured biological signal." OpenScientist did not deduce that labels had been randomized, but recognized a lack of expected biological patterns (reduced differential expression signal and inconsistent biomarker progression). This addresses a vulnerability in AI-driven science: the "garbage in, garbage out" problem. Rather than generating meaningless results from compromised data, OpenScientist exhibited epistemic humility, acknowledging limitations and qualifying conclusions accordingly. It noted that "biomarker progression trends showed opposite directions between datasets" and recommended experiments including "re-analyze Dataset R with alternative normalization to recover biological signal."

Case Study 5: Prespecified proteomic analysis of suvorexant clinical trial with blinded human-versus-AI adjudication

While the preceding case studies evaluated OpenScientist output post hoc across the original 66 runs (Fig. 2), we also directly compared independent analyses by a human team versus OpenScientist in a prespecified, preregistered comparison plan (Methods). We used a dataset from a randomized, placebo-controlled trial of the dual orexin receptor antagonist suvorexant in cognitively unimpaired adults, which previously showed that suvorexant acutely lowered CSF phosphorylated tau and amyloid-β.[39] The human analysis consisted of an in-progress manuscript draft (since submitted as a preprint),[40] which was compared to the OpenScientist report.

A domain expert adjudicated the blinded reports. The expert catalogued 19 findings across the reports: 3 shared, 13 unique to the human analysis, and 3 unique to OpenScientist. All were judged to be factually supported and all but one finding in the human analysis adhered to the prespecification. Secondary LLM adjudication differed in the number and granularity of findings identified, but when grouped by content the human expert and LLM comparisons were largely consistent (Extended Data Table 3).

The reports agreed on the primary endpoint: no protein was significantly differentially expressed after FDR correction in either CSF or plasma at either timepoint. They diverged primarily in how they interpreted secondary signals after this null result. The human analysis used nominally significant proteins for pathway and network enrichment and integrated those pathway patterns across time and compartment into a biological model. OpenScientist first tested for evidence of aggregate non-null signal, assessed reproducibility across time and compartments, and emphasized pathway findings robust to multiple thresholds or threshold-free analyses.

The reports converged on plasma at hour 24, where both describe proteostasis and vesicle-trafficking biology and where OpenScientist's distributional test found the strongest aggregate signal (1,277 nominally significant proteins against 538 expected). They differed at hour 16, where the human analysis reported 33 enriched terms in CSF and 22 in plasma and used them to describe an early central response, while OpenScientist

ACCELERATED ARTICLE PREVIEW

found fewer nominally significant proteins than chance would produce and confined its interpretation to hour 24. In summary, both analyses agreed on the primary result and differed in how each proceeded from it. The human analysis interpreted the full set of pathway-level results and assembled them into a mechanistic model, while OpenScientist restricted interpretation to signals it could show were reproducible and tested candidate mechanisms explicitly rather than inferring them. The two approaches are complementary, one oriented toward generating mechanistic hypotheses and the other toward establishing which signals are robust enough to support them.

Discussion

We present OpenScientist, an open-source agentic AI platform for semi-autonomous scientific analysis through transparent, iterative analysis, and reasoning. Across multiple clinical and biomedical use cases, OpenScientist integrated user-provided data with published literature, performed bespoke computational analyses, and synthesized mechanistic insights into coherent, verifiable results and testable hypotheses. These findings suggest that auditable, modular, and interpretable agentic AI can augment scientific workflows, particularly clinical studies requiring rapid integration of heterogeneous human datasets under constraints of time and expertise.

Across the case studies, OpenScientist reached conclusions broadly consistent with independent human analyses. In the prespecified human-vs-AI comparison, the primary conclusions were concordant despite different secondary analytical pathways. Direct comparison was complicated by the intended outputs: the human team was developing a manuscript requiring prioritization and synthesis within the constraints of scientific publishing, whereas OpenScientist produced a more comprehensive report, which also complicated blinding. Yet this divergence illustrates an opportunity for agentic AI. Scientific datasets permit many reasonable analytical trajectories, or a “garden of forking paths,” and AI makes it feasible to explore these alternatives and identify conclusions robust across analytical choices.[41] Further, the cost of this exploration is favorable: the median cost per run across the runs described in this study was $11.30 (range $0.72 to $117.07, Supp. Table 1).

A central contribution of OpenScientist is its open-source, self-hostable, and fully auditable architecture, which addresses key limitations of existing AI scientist platforms. Recently published systems vary substantially in openness: Google Co-scientist[3] is proprietary and closed source; Robin[7] releases its orchestration code but depends on proprietary Edison services for its full workflow; Biomni[8] provides an open-source release, although that release differs from the current hosted platform. In contrast, OpenScientist is freely available under a permissive license, can be installed on user-controlled infrastructure, and exposes its orchestration, analyses, reports, and execution traces, enabling researchers to inspect intermediate steps, validate outputs, and modify workflows. This transparency is complemented by evaluation strategies to reduce selective reporting, including reporting the complete denominator of attempted runs and failures, prespecified analyses, repeated runs, hidden evaluation, external validation, and randomized negative controls. Together, this framework provides a more realistic estimate of what current co-scientist systems can reliably accomplish.

OpenScientist’s modular skills architecture further enhances extensibility. By separating domain-agnostic reasoning from domain-specific analytical skills, researchers can extend functionality without altering core logic, supporting community-driven development of specialized skills for particular datasets, experimental paradigms, or biomedical domains.

OpenScientist demonstrated strong but imperfect performance. We noted several failure modes, including **(i) statistical reliability problems** (e.g., some missing multiple-testing correction across iterations, and residual false positives in negative-control and randomized analyses), **(ii) domain-science errors** in some skills (e.g., incorrect biochemical interpretation of metabolite ratios as flux proxies or inappropriate statistical treatment of

ACCELERATED ARTICLE PREVIEW

gene expression data), **(iii) architectural limitations** in the reporting and literature review pipeline that can compound summarization errors (e.g., the final report being generated from summaries rather than directly interrogating the data, and literature summaries lacking transparent citation tracking), and **(iv) reliance on latent knowledge instead of data** (Case Study 6).

Potential remediation approaches include updating Agent Skills (which in Case Study 3 remediated an observed failure; Supplementary Information, "Benchmarking OpenScientist with ad hoc updated skills"), explicit statistical safeguards, consistency checks across skills and scoring formulas, domain-aware validation layers for analysis code, and restructuring the report and literature synthesis stages so the agent can directly access data, provenance, and cited sources when generating conclusions. Continued refinement is likely to improve the expertise, accuracy, and impact of the system's analyses. As foundation models become more capable, the optimal degree of skill specification is likely to shift.

Current scientific AI systems, including OpenScientist, should be viewed as tools that support hypothesis testing and analysis rather than independent scientific actors. Expert oversight remains essential, particularly for mechanistic interpretation and causal inference. This aligns with Friedman's Fundamental Theorem of Biomedical Informatics and its extension to the AI era, which emphasize human-computer partnership.[42,43] Rather than replacing human scientists, OpenScientist can expand the breadth of analytical exploration while researchers provide contextual interpretation, prioritization, and scientific synthesis. A primary motivation for publishing at this stage is to bring the system to a wider community whose feedback and stress-testing will be essential to its improvement, consistent with the C-BRAIN mission of developing AI "by and for biomedical researchers".

We identify three immediate priorities for advancing autonomous scientific AI. First, validation frameworks must be strengthened: scientific AI reports should meet the same standards as scientific reports, including controlling for multiple comparisons with full accounting of all *in silico* experiments and pre-specified vs. post-hoc findings clearly identified, full disclosure of both positive and negative findings, and development and testing by expert scientists. Second, autonomous reasoning systems should be integrated with experimental validation, such as automated design of perturbation or intervention experiments to test causal claims. Third, prospective clinical and translational studies are needed to determine whether AI-generated predictions improve real-world patient outcomes. Only through such evaluation can the field determine whether platforms like OpenScientist represent transformative advances in biomedical discovery or sophisticated pattern-recognition systems that still require human supervision.

For sensitive data such as clinical information or unpublished analyses, OpenScientist supports an air-gapped mode in which an entire analysis runs end-to-end with no possibility of unintended data egress (Methods). Our analysis suggests that air-gapped mode yields similar analyses and conclusions to the online version (Extended Data Table 2; Supplementary Information, "Benchmarking OpenScientist with open weight models and in secure air-gapped mode") and that frontier open weight models yield similar scientific analysis and conclusions compared to closed-weight frontier models.

Taken together, these results support OpenScientist as a useful and still maturing co-scientist. Its open-source design enables deployment at any institution, allowing sensitive clinical and genomic data to remain within existing governance frameworks while supporting rigorous benchmarking and adversarial testing. By making both successes and limitations explicit, OpenScientist serves not only as a functional analysis platform but also as a testbed for defining standards of validation, transparency, and responsibility in autonomous scientific research. Because the use cases presented here all involved human clinical datasets, OpenScientist may be particularly well positioned to support clinical studies by analyzing complex data to yield interpretable findings, candidate biomarkers, and testable translational hypotheses. OpenScientist and other co-scientist approaches have enabled augmented research across the scientific process, and offer the future promise to deliver validated discoveries that could impact human health. Continued development focused on validation rigor,

ACCELERATED ARTICLE PREVIEW

experimental grounding, and prospective outcome testing will be essential for deployment in high-stakes biomedical settings.

Acknowledgements

This work was organized through and supported by C-BRAIN (c-brain.org). This work was also supported by resources and effort provided by the Tracy Family Stable Isotope Labeling Quantitation Center (Principal Investigator (PI) RJB) established by the Tracy Family, Richard Frimel & Gary Werths, GHR Foundation, Pat and Jane Tracy, Anonymous, Anne & Ray Capestrain, Community Foundation Serving West Central Illinois and Northeast Missouri, JTL Family Fund, Payne Family, Mary & Jay Sullivan, Tracy Family Foundation, Catherine & Tom Tracy, Community Foundation for the Land of Lincoln, Jim & Jil Tracy, Joe & Jill Tracy, Sonja & Robert M. Willman, Boniface Foundation, Jean & Michael Buckley, Ann Liberman, Clemence S. Lieber Foundation, Mary Schoolman & Dr. James Hinrichs, and Susan & Scott Stamerjohn brought together by The Foundation for Barnes-Jewish Hospital. The results published here are in part based on data obtained from the NF Data Portal (http://www.nf.synapse.org, RRID:SCR_021683) and made available through the NF Open Science Initiative. We thank Drs. Suzanne Schindler, Siobhan Sutcliffe, Ganesh Babulal, and Eric Lenze for their guidance of the SEABIRD study. Figure illustrations were created by Karen Thiebes, Ph.D. (Simplified Science Publishing, LLC).

Funding Statement

This work was supported by C-BRAIN (c-brain.org). SEABIRD was funded through NIA RF1AG061900 (PI: R.J.B.). J.H.C., N.L.H., C.J.M. and J.T.R. were supported by the NIH Monarch Initiative 2R24OD011883 and by the Director, Office of Science, Office of Basic Energy Sciences of the U.S. Department of Energy Contract No. DE-AC02-05CH11231. The suvorexant study was funded by the Washington University Personalized Medicine Initiative and Good Ventures.

Author Contributions Statement

J.T.R., R.J.B., C.J.M., E.C.L., K.F.R. and Z.B.A. conceived and designed the study. J.T.R., L.C., J.H.C., and C.J.M. developed and deployed the OpenScientist software. K.F.R., Z.B.A., J.T.R., R.J.B., K.Y., M.M., M.B., Y.L., J.B., D.G., J.A.M., and E.S.M. ran and evaluated OpenScientist analyses. R.J.B., T.L.S.B., S.F., B.A.G., N.R.B., J.G.B., Y.H., V.O., M.L., Y.L., M.M., K.Y., M.B., B.P.L., T.P., J.Y., J.B., L.F., D.L.E., M.P., D.G. and J.A.M. provided data and resources. K.F.R., K.Y., M.B., and J.T.R. prepared the figures. E.C.L., S.A.N., A.A.O., N.H., S.B., and N.L.H. provided project administration and coordination. K.F.R., Z.B.A., J.T.R. and R.J.B. wrote the manuscript with input from all authors; J.T.R., R.J.B. and C.J.M. supervised the work and acquired funding. All authors reviewed and approved the final manuscript.

Competing Interests Statement

R.J.B. has received research funding from Avid Radiopharmaceuticals, Janssen, Roche/Genentech, Eli Lilly, Eisai, Biogen, AbbVie, Bristol-Myers Squibb, and Novartis. Washington University and R.J.B. have equity ownership interest in C2N Diagnostics. Washington University, R.J.B., V.O., J.G.B., and N.R.B. receive income based on technology (stable isotope labeling kinetics, blood plasma assay, methods of diagnosing AD with phosphorylation changes, neurofilament light chain assays and materials) licensed by Washington University

ACCELERATED ARTICLE PREVIEW

to C2N Diagnostics. R.J.B. receives income from C2N Diagnostics for serving on the scientific advisory board. R.J.B. serves as an unpaid member on a scientific advisory board for Biogen. T.L.S.B. has served as a paid and unpaid consultant to Eisai, Siemens, Biogen, Janssen, Eli Lilly, Sora Neuroscience and Bristol-Myers Squibb. B.P.L. receives consulting fees from Eli Lilly, Takeda, Vesalius, and the Weston Family Foundation. B.P.L. serves on Data Safety and Monitoring Boards for Eli Lilly. B.P.L. serves on the Scientific Advisory Board for Beacon Biosignals and receives compensation as a scientific advisor to Applied Cognition. B.P.L. receives drug/matched placebo from Merck for a clinical trial funded by a private foundation and drug/matched placebo from Eisai for a clinical trial funded by the NIA. E.S.M. is a scientific advisor for Takeda Pharmaceuticals and Eli Lilly, and has received research funding from Eisai. All other authors declare no competing interests.

Tables

Case Study	Purpose	Query	Data	Additional materials	Scientific output
1. SEABIRD: Alzheimer's blood biomarkers in a community cohort	Data analysis	Read SEABIRD grant scientific plan and carry out specified analyses.	CSV including amyloid PET, blood biomarkers, and demographics by participant (n=325)	Grant proposal, data dictionaries	Analyzed biomarker performance demonstrating utility of %pTau217 for PET status prediction.
2. Survival prediction by human plasma proteomics	Supervised survival modeling	Build a survival prediction model maximizing concordance index on held-out testing set.	CSV of plasma proteomics and demographic data (n=500)	None	Generated survival model utilizing proteomics with improved concordance on held out data and identified most impactful proteins.
3. Neurofibrillary tangle (NFT) transcriptomics in human brain to investigate lysosomal acidification	Hypothesis investigation	Investigate how proteostasis, in particular lysosomal acidification, is rewired by tau pathology.	h5ad file of single cell transcriptomics of neurons with and without NFTs	None	Identified altered expression of transcripts significant to lysosomal acidification in NFTs.
4. Multiple myeloma (MM) transcriptomics for hypothesis generation and validation	Hypothesis generation	Identify differentially expressed genes between clinical subgroups and hypothesize a biological mechanism for progression of myeloma.	Microarray data with patients stratified by clinical subgroup (MGUS, SMM, MM) (n=99)	None	Hypothesized UPR failure mechanisms are central to MM progression.
	Hypothesis validation with negative control	Given a set of hypotheses, validate on two datasets.	External microarray validation dataset (n=162), scrambled copy	OpenScientist output from initial hypothesis generating run	Appropriately rejected hypotheses in randomized data.
5. Prespecified proteomic analysis of suvorexant clinical trial	Prespecified analysis with blinded human-vs-AI adjudication	Follow the preregistered analysis plan and carry out specified analyses.	CSV of SOMAscan proteomic measurements from CSF and plasma at two timepoints in suvorexant 20 mg (n=12) and placebo	Pre-registered analysis plan	Identified no individual proteins significant after FDR correction, while highlighting a robust plasma proteostasis and vesicle-trafficking signal at hour 24.

ACCELERATED ARTICLE PREVIEW

			(n=13) groups		

Table 1. *Summary of OpenScientist case studies featured in this manuscript.*

Figure Legends

Figure 1*. Overview of the OpenScientist scientific investigation workflow. 1) Scientist describes their scientific task (a query, a scientific analysis, etc.), and optionally includes one or more data files to support the task. 2) The agent is given the task and data files, and autonomously investigates by querying PubMed, examining papers, writing and executing code, guided by a suite of Agent Skills, comprising advice in various scientific domains for the agent. 3) The artifacts from the investigation, including finding, hypotheses, analyses, plots, are saved in a knowledge state data structure (KSDS). 4) Optionally (if the user selected co-investigate mode), feedback from the user is collected and stored in the KSDS. The new iteration of OpenScientist then begins (step 2), with the KSDS provided to build on existing knowledge. 5) After N iterations (10 by default), the results are compiled into a final report.*

Figure 2*: CONSORT diagram outlining all OpenScientist runs during the development period. Two runs failed due to user error. Three runs had technical failures in OpenScientist. Of the remaining 88 OpenScientist runs, 22 did not include a dataset and were limited to a literature search query. In order to emphasize the analytic capability of OpenScientist beyond literature search alone, these were excluded from contention as featured case studies. All 66 remaining runs ran to completion and were split into 28 query types, some of which were run multiple times, and covered 13 broad data types. Of the 28 possible query types, 7 are included as case studies, with user engagement and completion of expert evaluation of OpenScientist results being the primary drivers for inclusion. Supplementary materials for the case studies (including input files, where possible, and all OpenScientist outputs) are available on Zenodo (https://doi.org/10.5281/zenodo.18852838)*

Figure 3: a) *Workflow: Case Study 1 utilized the SEABIRD dataset which included plasma biomarkers of Alzheimer's disease, PET Centiloid values, cognitive scores, and patient demographics from a community cohort. OpenScientist was provided the SEABIRD grant proposal and dataset and instructed to carry out the specified analysis plan which included ROC analysis to evaluate the ability of plasma biomarkers for discriminating PET status.* ***b)*** *Comparison of human generated analysis of the dataset (left) and OpenScientist generated analysis (right) showing equivalent conclusions.*

Figure 4. a) *Workflow: Case Study 2 utilized a dataset including plasma proteomics and demographic data. OpenScientist was prompted to build a survival model.* ***b)*** *OpenScientist output graph comparing model performance and identifying proteomic features predictive of survival.* ***c)*** *OpenScientist shows comparable performance by c-index compared to published Biomarkers of Aging (BoA) models. † Internal validation (50/50 train-test split); not evaluated on competition holdout set.*

Figure 5: a) *Workflow: Data included single-soma transcriptomics from neurons with and without neurofibrillary tangles (NFTs). OpenScientist was instructed to investigate the relationship between proteostasis and lysosomal acidification.* ***b-c)*** *OpenScientist analyzed gene expression changes and generated a hypothesis about a relationship between neurons with NFTs and impaired lysosomal acidification via downregulation of lysosomal ion channels that was verified by a human expert.*

ACCELERATED ARTICLE PREVIEW

References

1. Lam, R. *et al.* Learning skillful medium-range global weather forecasting. *Science* **382**, 1416–1421 (2023).
2. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
3. Gottweis, J. *et al.* Accelerating scientific discovery with Co-Scientist. *Nature* **655**, 487–496 (2026).
4. Lu, C. *et al.* The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. Preprint at https://doi.org/10.48550/arXiv.2408.06292 (2024).
5. Rosalind. https://openai.com/rosalind/.
6. Mitchener, L. *et al.* Kosmos: An AI Scientist for Autonomous Discovery. Preprint at https://doi.org/10.48550/arXiv.2511.02824 (2025).
7. Ghareeb, A. E. *et al.* A multi-agent system for automating scientific discovery. *Nature* **655**, 497–505 (2026).
8. Huang, K. *et al.* Autonomous biomedical research with an artificial intelligence agent. *Science* **393**, eadz4351 (2026).
9. Li, O., Agarwal, V., Zhou, S., Gopinath, A. & Kassis, T. K-Dense Analyst: Towards Fully Automated Scientific Analysis. Preprint at https://doi.org/10.48550/arXiv.2508.07043 (2025).
10. Claude Science. https://www.anthropic.com/news/claude-science-ai-workbench.
11. Gao, S. *et al.* Empowering biomedical discovery with AI agents. *Cell* **187**, 6125–6151 (2024).
12. Wei, J. *et al.* From AI for Science to Agentic Science: A Survey on Autonomous Scientific Discovery. Preprint at https://doi.org/10.48550/arXiv.2508.14111 (2025).
13. Roberts, K. F. *et al.* Towards an AI biomedical scientist: Accelerating discoveries in neurodegenerative disease. *The Journal of Prevention of Alzheimer's Disease* **13**, 100398 (2026).
14. Miao, J., Davis, J. R., Zhang, Y., Pritchard, J. K. & Zou, J. Reimagining research papers as interactive and reliable AI agents. *Nature* 1–9 (2026) doi:10.1038/s41586-026-11044-y.

ACCELERATED ARTICLE PREVIEW

15. Li, M. *et al.* An Alzheimer's disease blood test study in a diverse community-based population: Preliminary results from SEABIRD. *Alzheimer's & Dementia* **19**, e073358 (2023).
16. Bateman, R. *et al.* Alzheimer's disease blood tests of amyloid- beta 42/40, %p- tau217, 181, and 205 ratios and MTBR- 243 in real- world populations: Results from SEABIRD and BioFINDER2. *Alzheimer's & Dementia* **20**, (2025).
17. Ying, K. *et al.* An open competition for biomarkers of aging. *Nat Aging* **6**, 1193–1195 (2026).
18. Beydoun, M. A. *et al.* Plasma neurofilament light and its association with all-cause mortality risk among urban middle-aged men and women. *BMC Med* **20**, 218 (2022).
19. Chen, X. *et al.* CX3C chemokine: Hallmarks of fibrosis and ageing. *Pharmacological Research* **208**, 107348 (2024).
20. Moqri, M. *et al.* Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell* **186**, 3758–3775 (2023).
21. Ying, K. *et al.* A unified framework for systematic curation and evaluation of aging biomarkers. *Nat Aging* **5**, 2323–2339 (2025).
22. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11**, 303–327 (2019).
23. Lu, A. T. *et al.* DNA methylation GrimAge version 2. *Aging* **14**, 9484–9549 (2022).
24. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging* **10**, 573–591 (2018).
25. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell* **49**, 359–367 (2013).
26. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol* **14**, R115 (2013).
27. Belsky, D. W. *et al.* DunedinPACE, a DNA methylation biomarker of the pace of aging. *eLife* **11**, e73420 (2022).
28. Sugden, K. *et al.* Association of Pace of Aging Measured by Blood-Based DNA Methylation With Age-Related Cognitive Impairment and Dementia. *Neurology* **99**, e1402–e1413 (2022).
29. Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med* **11**, 54 (2019).

ACCELERATED ARTICLE PREVIEW

30. McEwen, L. M. *et al.* The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells. *Proceedings of the National Academy of Sciences* **117**, 23329–23335 (2020).
31. Ying, K. *et al.* Causality-Enriched Epigenetic Age Uncouples Damage and Adaptation. *Nat Aging* **4**, 231–246 (2024).
32. Horvath, S. *et al.* Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging (Albany NY)* **10**, 1758–1775 (2018).
33. Lin, Q. *et al.* DNA methylation levels at individual age-associated CpG sites can be indicative for life expectancy. *Aging (Albany NY)* **8**, 394–401 (2016).
34. Otero-Garcia, M. *et al.* Molecular signatures underlying neurofibrillary tangle susceptibility in Alzheimer's disease. *Neuron* **110**, 2929-2948.e8 (2022).
35. Consortium, T. P. *et al.* A Comprehensive Enumeration of the Human Proteostasis Network. 1. Components of Translation, Protein Folding, and Organelle-Specific Systems. 2022.08.30.505920 Preprint at https://doi.org/10.1101/2022.08.30.505920 (2022).
36. Consortium, T. P. *et al.* A Comprehensive Enumeration of the Human Proteostasis Network. 2. Components of the Autophagy-Lysosome Pathway. 2023.03.22.533675 Preprint at https://doi.org/10.1101/2023.03.22.533675 (2023).
37. Li, P., Gu, M. & Xu, H. Lysosomal Ion Channels as Decoders of Cellular Signals. *Trends Biochem Sci* **44**, 110–124 (2019).
38. Lopez-Corral, L. *et al.* Transcriptome analysis reveals molecular profiles associated with evolving steps of monoclonal gammopathies. *Haematologica* **99**, 1365–1372 (2014).
39. Lucey, B. P. *et al.* Suvorexant Acutely Decreases Tau Phosphorylation and Aβ in the Human CNS. *Ann Neurol* **94**, 27–40 (2023).
40. Pedersen, T. J. *et al.* Suvorexant Inversely Affects Phosphorylated Tau in Human Cerebrospinal Fluid and Plasma. 2026.11.03.26367196 Preprint at https://doi.org/10.64898/2026.11.03.26367196 (2026).
41. Miao, J., Pritchard, J. K. & Zou, J. The Agentic Garden of Forking Paths. Preprint at https://doi.org/10.48550/arXiv.2607.01507 (2026).
42. Friedman, C. P. A "Fundamental Theorem" of Biomedical Informatics. *J Am Med Inform Assoc* **16**, 169–170 (2009).

ACCELERATED ARTICLE PREVIEW

43. Payne, P. R. O., Chen, J. H. & Longhurst, C. A. Extending the Fundamental Theorem of Biomedical Informatics for the AI era. *J Am Med Inform Assoc* **33**, 1415–1419 (2026).

Methods

System Design and Architecture

OpenScientist operates as an analysis loop that iteratively refines hypotheses through computational analysis and literature integration (Fig. 1). OpenScientist uses an AI agent (Claude Code,[44] codex (CLI),[45] and omp[46] are currently supported), an LLM (models may be accessed from GCP Vertex AI, Azure / Microsoft Foundry, AWS Bedrock, or any Anthropic- or OpenAI-compatible API), and a public collection of Agent Skills (https://github.com/K-Dense-AI/claude-scientific-skills)[47] for specific biomedical knowledge domains, analytical functions, or bioinformatics tools. Supported providers include Anthropic, Google Vertex AI, AWS Bedrock, Azure Foundry, and CBORG (an internal compute resource at Lawrence Berkeley National Laboratory). Domain-specific tools for data analysis, literature search, and knowledge synthesis are exposed to the agent via Agent Skills and the Model Context Protocol (MCP). We investigated performance of OpenScientist using agents including Claude Code[44] and omp,[46] and LLMs including Sonnet 4.5,[48] Opus 4.5,[49] Opus 4.6,[50] Opus 4.8,[51] Kimi K3,[52] and GLM 5.2[53] in this study (model versions provided in Supp. Table 1). Supplementary runs were accessed via Azure Microsoft Foundry using Opus 4.6 (claude-opus-4-6, model version 1), Opus 4.8 (claude-opus-4-8, model version 1), Kimi K3 (FW-Kimi-K3 version 1) and GLM 5.2 (FW-GLM-5.2 version 1). The Agent Skills encapsulate capabilities such as statistical analysis, literature retrieval, knowledge graph reasoning, and biomedical interpretation. The system runs within a Docker container for security, portability, and reproducibility, and is orchestrated by an agent that dynamically selects and combines skills based on the user's query and the evolving state of the investigation. Code generated by the agent was executed in a containerised Python environment. For the featured case studies (runs between November 2025 and February 2026) this was Python 3.11 with the PyPI releases current when the hosted instance was built (pandas 2.3, numpy 2.4, scipy 1.16 to 1.17, scikit-learn 1.8, statsmodels 0.14.6, matplotlib 3.10), orchestrated by Claude Code 2.0.x (2.0.37 for Case Study 1); for the supplementary experiments (August to October 2026) it was Python 3.12.13 with package versions pinned per commit in the repository (pandas 3.0.0, numpy 2.3.5, scipy 1.16.3, scikit-learn 1.8.0, statsmodels 0.14.6, scanpy 1.12, anndata 0.12.6, matplotlib 3.10.8, seaborn 0.13.2; Claude Code 2.1.228, omp 17.1.5).

Scientists initiate an analysis by entering a free-text query describing the scientific task, optionally including one or more datasets to guide the investigation, as well as any supplementary reference material (articles, slide decks, etc). Once initialized, OpenScientist autonomously executes an iterative analysis loop. The system performs N iterations of research cycles, each comprising a set of autonomous research actions by the agent. By default N is 10, though the user can increase the iteration count for more complex analyses. Early stopping is available to the agent (although this was not used in any run for any case study analyzed in this study). In each iteration, OpenScientist evaluates the query, reviews findings and outputs from any prior iterations, selects which hypotheses to pursue (a judgment made by the agent in light of the research question, the data available and the current analysis and findings), executes new analyses on user-provided datasets, performs targeted searches of the scientific literature, and retrieves and reads relevant abstracts. It then integrates new

ACCELERATED ARTICLE PREVIEW

and prior evidence into an updated set of findings, which are saved in a knowledge state data structure (KSDS) in the form of a JSON file. This iterative loop allows OpenScientist to refine its reasoning, expand its evidence base, and converge toward coherent, evidence-supported conclusions.

At the end of the run, OpenScientist uses the KSDS to produce a final report summarizing all analyses, findings, synthesized knowledge, and evidence-based conclusions. It includes potential follow-up experiments or analyses, when appropriate.

OpenScientist also supports a secure, completely disconnected air-gapped mode that guarantees that no unintended data egress is possible. The agent accesses a local copy of PubMed for literature searching, prohibits tools from connecting to the internet, and performs LLM inference from a locally hosted model (or optionally, from a remote API, as an explicit exception). This was used in benchmarking secure air-gapped mode in Case Study 3 (Supplementary Information). Unless otherwise specified, air-gapped mode was not used in this study.

Case studies

For Case Study 1, the dataset was the subset of SEABIRD participants with blood collection and confirmatory amyloid PET imaging.[15] Plasma biomarkers (including Aβ42, Aβ40, pTau181, pTau205, pTau217, and total tau) were enriched via immunoprecipitation and measured using liquid chromatography-tandem mass spectrometry.[54,55] Accompanying data included demographic variables (age, sex, race), APOE genotype, cognitive and functional assessments (AD8, Montreal Cognitive Assessment [MoCA], and Clinical Dementia Rating sum of boxes [CDR-SB]). Amyloid PET Centiloid values provide a quantitative reference standard for cerebral amyloid burden.[56] The prompt for OpenScientist included the raw data in tabular format, data dictionaries, and the grant application for the study, which included relevant background information, the study design, aims, and the analytical plans; the query instructed OpenScientist to use the proposal to conduct a full analysis of the data based on the aims and goals of the study. The human-led comparison analyses were performed in Statistical Analysis System (SAS). Across the ten exploratory runs, issues were resolved by explicitly specifying missing-data handling rules, required reporting standards, cohort boundaries, and the MRI-free Centiloid cutoff of 18.41 for amyloid positivity, as well as by uploading a deduplicated dataset.[57]

For Case Study 2, proteomic measurements were generated using two Alamar NULISA panels targeting central nervous system and inflammation-related proteins. Accompanying metadata includes demographic variables such as age and sex, as well as survival outcomes defined by time to death. The instructions specified a reproducible 50:50 training:testing split, inclusion of all proteomic and metadata features, and comparison against a baseline model restricted to age and sex. The query allowed flexibility in model choice, including traditional Cox proportional hazards models and modern machine learning approaches such as penalized Cox or random survival forests, with explicit encouragement to tune or regularize models to improve performance. For the strict Case Study 2 rerun, we deduplicated the official Phase 2 proteomic subset by Plasma.ID and used a mortality-stratified 50:50 split (seed 42): 175 training participants (16 deaths) and 176 hidden evaluation participants (17 deaths). OpenScientist received the training proteomics and survival labels, with age available only as optional auxiliary supervision, and did not receive the hidden evaluation proteomics, labels, phenotypes, competition submissions, leaderboard, or challenge results. The returned six-protein molecular score was frozen before local evaluation; its molecular-age weight was 0. Local scoring on the hidden evaluation proteomics yielded a c-index of 0.826.

For Case Study 3, the dataset was obtained from CellxGene and subsetted to two populations of excitatory neurons (Ex1 and Ex2) for analysis (exploratory runs: Supp. Table 4).[34] For the reproducibility analysis, we repeated the same query and dataset across three experiments. First, 10 independent runs using Claude Code

ACCELERATED ARTICLE PREVIEW

with Claude Opus 4.6 were assessed for agreement on the main findings, gene selection, and reliance on statistically significant genes with small effect sizes. Second, we compared Claude Code with Opus 4.8 and omp with Kimi K3 or GLM 5.2 across online and network-isolated versions (10 runs per configuration and version; 60 runs). We evaluated agreement on biological conclusions, gene selection, effect directions, and donor-level consistency. Third, we compared the original and revised genomics skills (15 runs per version; 30 runs). Reports and execution records were assessed for recognition of pseudoreplication, donor-aware inference, and execution of raw-count pseudobulk analysis with donor mixed-model confirmation and reconciliation of results. We additionally compared findings from iteration 1 with those from all iterations across the ten Opus 4.6 runs, and repeated the query with the KSDS ablated (Supp. Tables 5 and 6).

For Case Study 4, hypotheses were generated from the López-Corral et al. (2014) microarray dataset (GSE47552; n=99 samples, 33,297 genes)[38] and validated using an external cohort (Mayo Clinic, GSE6477, n=162; Dataset V) and a copy with randomized disease labels (Dataset R) (Supp. Table 7). Both contained samples spanning the complete disease spectrum from normal plasma cells through MGUS, SMM, newly diagnosed MM, and relapsed MM. Dataset R served as a negative control: we randomized disease stage labels while preserving all other data characteristics, effectively eliminating biological signal while maintaining technical structure. OpenScientist was tasked with independently validating the four hypotheses using both datasets without prior knowledge of the randomization manipulation. To estimate the false positive rate of the >50 DEG criterion in comparisons of different stages and evaluation of hypotheses, we regenerated 1000 more datasets with labels permuted as in Dataset R and recomputed DEGs as before, computing the FDR for DEGs, for stage comparisons, and for evaluations of H3 using this >50 DEG criterion.

For Case Study 5, the aspredicted.org preregistration template specifying primary question, dependent variables, planned analyses, and candidate biological domains of interest was completed by the human analysis team that generated the data and registered before AI analysis (AsPredicted #298545). We preregistered two datasets and only one moved forward to adjudication due to time constraints. OpenScientist received the preregistration and four differential expression tables (suvorexant 20 mg (n=12) and placebo (n=13) groups, CSF and plasma, hours 16 and 24 normalized to a common hour 0 baseline; logFC, nominal p-value, and Benjamini-Hochberg adjusted p-value for 10,751 SOMAscan aptamers). In parallel and independently, the human team that collected the data analyzed the same datasets. Both reports were converted to a common format, stripped of identifying features, and relabeled as Analysis A (human) and Analysis B (OpenScientist). An independent domain expert adjudicator received the preregistration, Analyses A and B, a structured adjudication table, and written instructions. The adjudicator enumerated each finding (defined as an independent idea containing data and interpretation) in the blinded reports and evaluated concordance, adherence to the prespecified plan, and factual validity. In addition to the adjudication table, the expert also provided narrative feedback on perceived utility of the tool. The primary adjudication was performed by an independent domain expert in sleep and neurodegeneration who took no part in either analysis. Secondary adjudication was performed by Claude Opus 5[58] and GPT-5.6 Sol[59] using identical instructions and materials. Methods for Case Studies 6–8 are given in the Supplementary Information (Supp. Tables 8–10).

Data Availability

For the featured main text and supplementary case studies, the query, the input data (where possible), and all outputs are available in the Supplementary Information and on Zenodo (https://doi.org/10.5281/zenodo.18852838). SEABIRD and the suvorexant study were approved by the Washington University in St. Louis IRB (SEABIRD, IRB 201902081, NCT03899844; suvorexant, IRB

ACCELERATED ARTICLE PREVIEW

201908156, NCT03077620) with written informed consent from all participants. Data can be obtained with approval under data use agreement from the corresponding author. All other datasets are public (CELLxGENE; GEO accessions GSE6477 and GSE47552; Harvard Dataverse doi:10.7910/DVN/8S7JVS; NF Data Portal, RRID:SCR_021683).

Code Availability

OpenScientist is accessible through a public web interface at openscientist.io, which is provided free of charge as resources allow. The source code is available under a permissive Apache 2.0 license at https://github.com/openscientist-io/openscientist for users who prefer to run OpenScientist on their own server infrastructure. For sensitive data, self-hosted deployments can additionally be configured in an air-gapped mode to remove the possibility of data egress (see System Design and Architecture). Analysis scripts used to generate the figures and tables are deposited with the case-study data on Zenodo (https://doi.org/10.5281/zenodo.18852838).

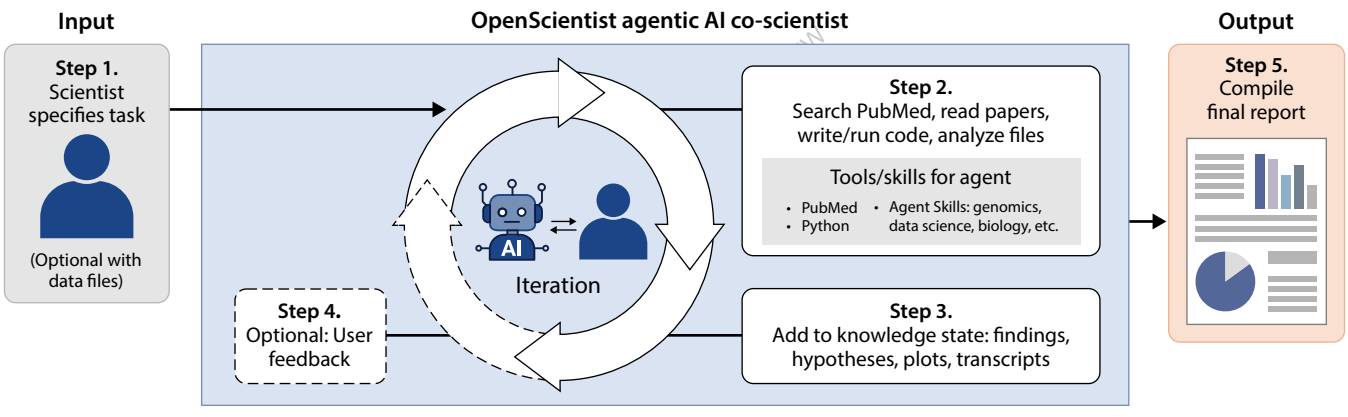
Methods-only References

44. Anthropic. Claude Code. https://github.com/anthropics/claude-code.
45. codex. https://chatgpt.com/codex/.
46. omp. https://omp.sh/.
47. K-Dense AI. scientific-agent-skills. https://github.com/K-Dense-AI/scientific-agent-skills.
48. Sonnet 4.5. https://www-cdn.anthropic.com/963373e433e489a87a10c823c52a0a013e9172dd/Claude%20Sonnet%204.5%20System%20Card.pdf.
49. Opus 4.5. https://www-cdn.anthropic.com/bf10f64990cfda0ba858290be7b8cc6317685f47/Claude%20Opus%204.5%20System%20Card.pdf.
50. Opus 4.6. https://assets.claude.com/6a5fa276ac68b9aeb0c8b6af5fa36326e0e166dd.pdf.
51. Opus 4.8. https://www-cdn.anthropic.com/0f0c97ad20d8005706296bd92aa1c27c6b2f4f61/Claude%20Opus%204.8%20System%20Card.pdf.
52. kimi k3. https://huggingface.co/moonshotai/Kimi-K3.
53. GLM 5.2. https://huggingface.co/zai-org/GLM-5.2.

ACCELERATED ARTICLE PREVIEW

54. Schindler, S. E. *et al.* High-precision plasma β-amyloid 42/40 predicts current and future brain amyloidosis. *Neurology* **93**, e1647–e1659 (2019).
55. Barthélemy, N. R. *et al.* Highly accurate blood test for Alzheimer's disease is similar or superior to clinical cerebrospinal fluid tests. *Nat Med* **30**, 1085–1095 (2024).
56. Klunk, W. E. *et al.* The Centiloid Project: standardizing quantitative amyloid plaque estimation by PET. *Alzheimers Dement* **11**, 1-15.e1–4 (2015).
57. Hoagey, D. A. *et al.* The Knight Alzheimer Research Imaging (KARI) dataset: a comprehensive multimodal resource for exploring aging, preclinical, and symptomatic Alzheimer disease pathology. *Res Sq* rs.3.rs-7962593 (2025) doi:10.21203/rs.3.rs-7962593/v1.
58. Opus 5. https://assets.claude.com/ceaf5c7ff2783855203fde8208ec311252dced5b.pdf.
59. GPT 5.6. https://openai.com/index/gpt-5-6/.

ACCELERATED ARTICLE PREVIEW


Input
OpenScientist agentic AI co-scientist
Output
Step 1.
Scientist specifies task
(Optional with data files)
Step 2.
Search PubMed, read papers, write/run code, analyze files
Tools/skills for agent
PubMed
Python
Agent Skills: genomics, data science, biology, etc.
Step 3.
Add to knowledge state: findings, hypotheses, plots, transcripts
Step 4.
Optional: User feedback
AI
Iteration
Step 5.
Compile final report

OpenScientist total runs n = 93
11/15/2025 - 2/16/2026

Excluded: user error, incomplete query or incorrect upload **n = 2**

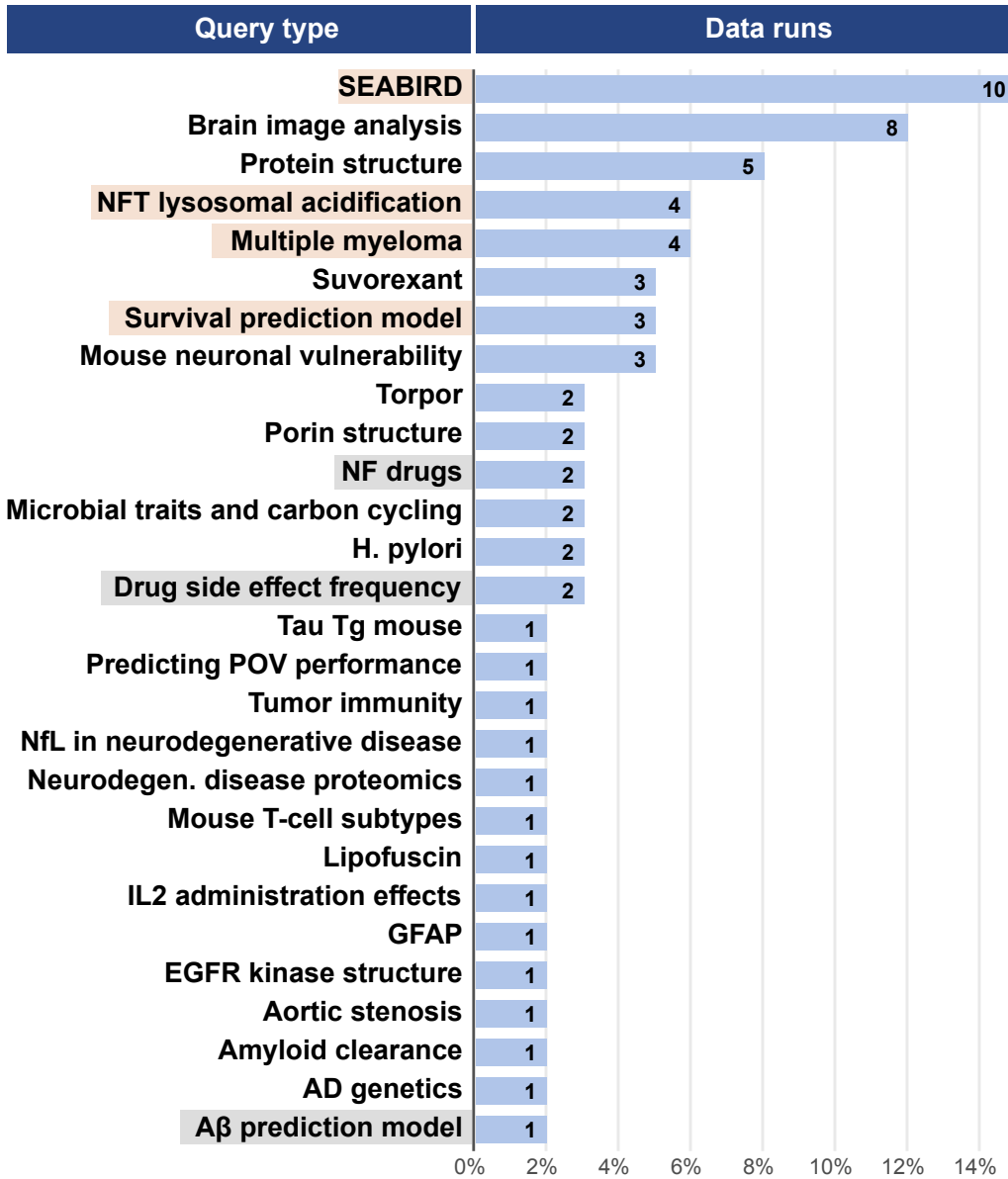
OpenScientist inputs n = 91

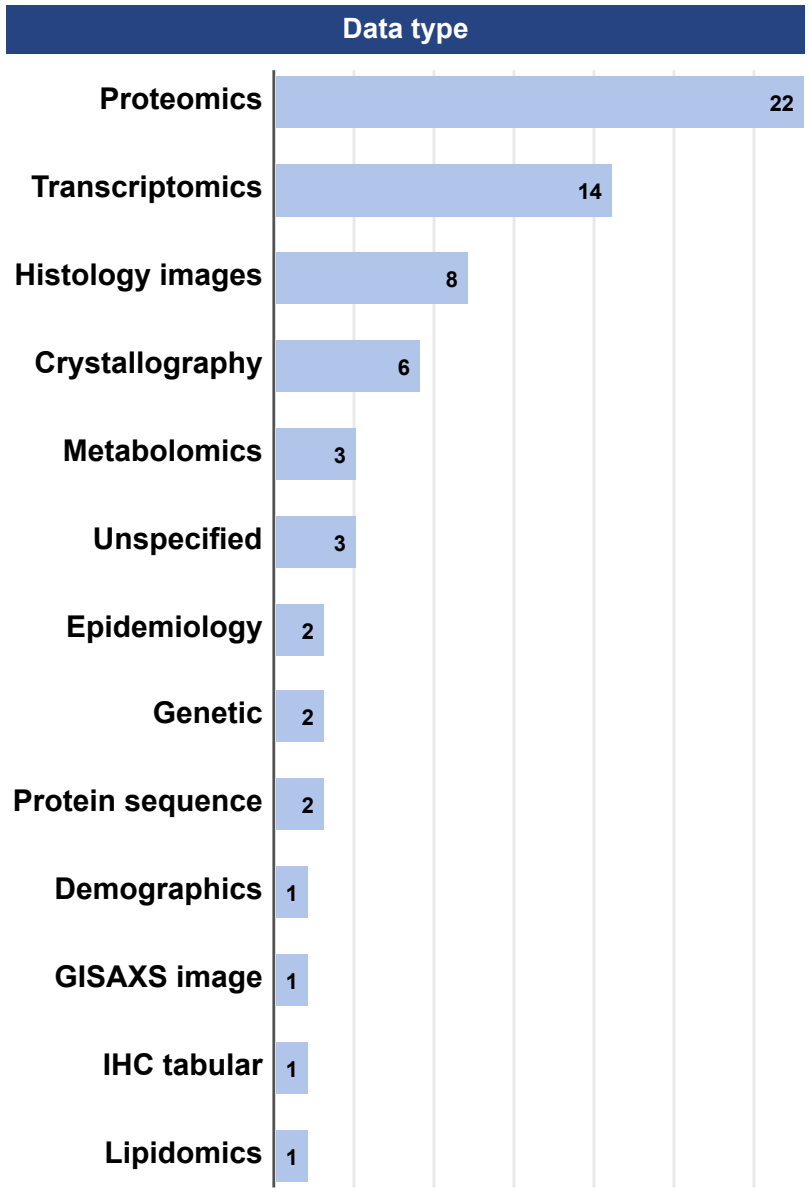
Excluded: failed run **n = 3**

OpenScientist completed runs n = 88

Excluded: no dataset included/literature search only **n = 22**

OpenScientist data runs n = 66





a

SEABIRD Grant Proposal
Prespecified analysis plan

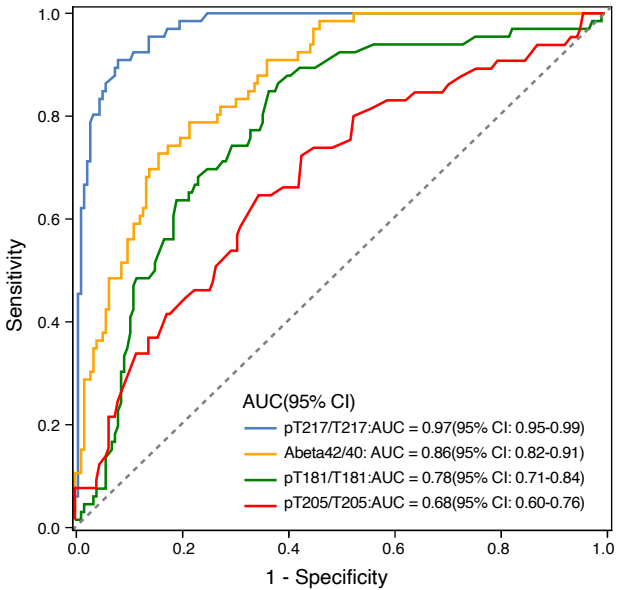
SEABIRD Dataset
Patient demographics, PET data, cognitive metrics, plasma biomarkers (n = 325)

OpenScientist AI Platform
Instructions:
Read the grant proposal and carry out the prespecified analysis plan

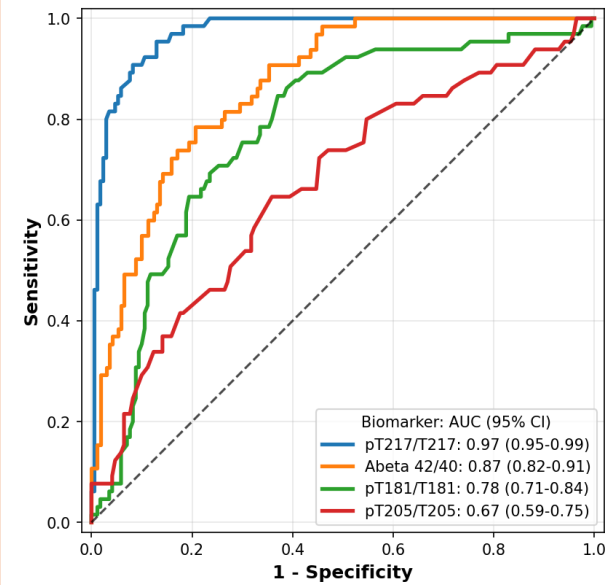
Output
Receiver operating characteristic (ROC) analysis to evaluate discriminative performance of biomarkers for PET status and clinical utility analysis of screening strategies

b

Human analyst-generated graph



OpenScientist-generated graph



a
Proteomic Dataset
Patient demographics and plasma proteomics (n = 500)
AI
OpenScientist AI Platform
Instructions:
Build a survival prediction model maximizing concordance index on a held-out testing set
Output
Model concordance index improvement over demographic only model, Kaplan Meier survival curves by risk group from model, most predictive proteomic features for survival
b
OpenScientist-generated graph
Model Performance Comparison
C-index
0.580
0.615
0.740
0.789
0.796
Age Only
Age+Sex
Simple Score (9 markers)
20-Marker Panel
Full Model (374 proteins)
c
AI Agent vs Published Biological Clocks
Published BoAs
AI Agent (OpenScientist)
Concordance Index (C-index)
C = 0.812
C = 0.805
C = 0.796
C = 0.791
C = 0.781
GrimAgeV2
GrimAgeV1
OpenScientist (P)-H
PhenoAge
Hannum
Zhang_10
Horvathv1
PEDBE
Horvathv2
YingCausAge
YingDamAge
Lin
YingAdaptAge
DunedinPACE
DunedinPoAm38

a

Neuron Transcriptome Dataset
Single neuron transcriptomic dataset

OpenScientist AI Platform

Instructions:
Investigate how proteostasis, in particular lysosomal acidification, is rewired by tau pathology

Output
Gene expression fold changes and statistics

b

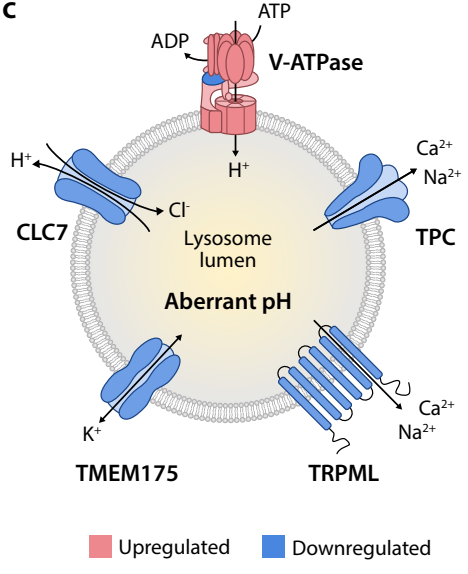
OpenScientist-generated table

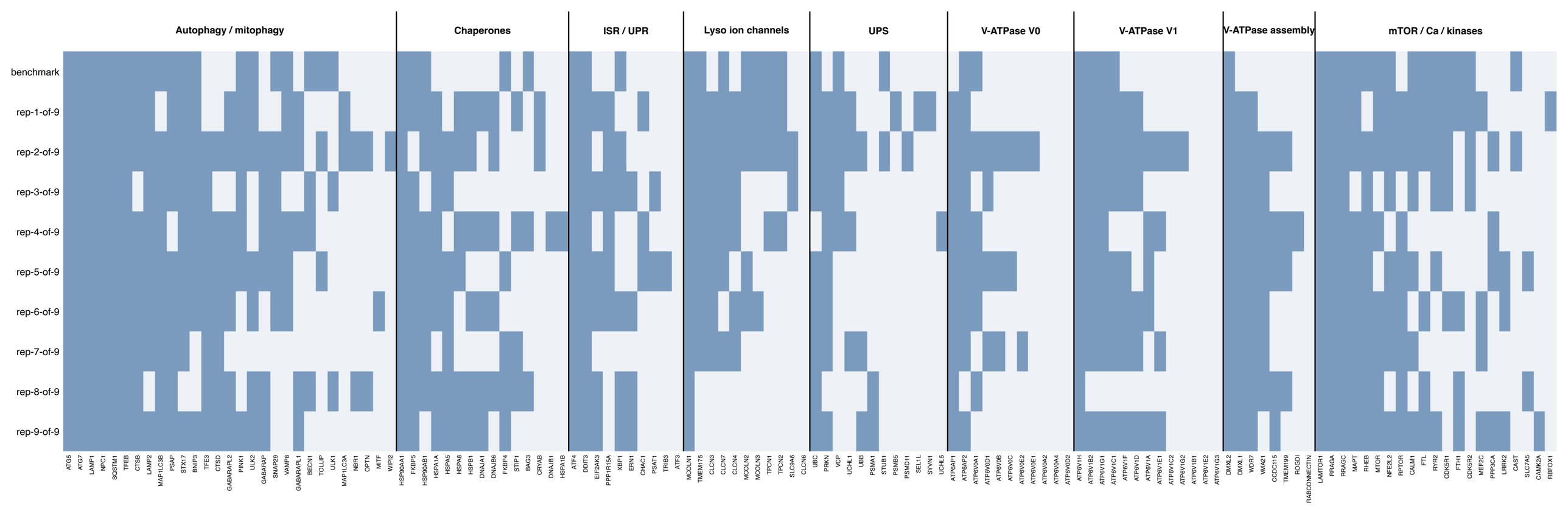
Component	Direction	Key Genes	Statistical Evidence
V-ATPase subunits	↑ Upregulated	ATP6V1B2, ATP6AP2	30/38 subunits up; ATP6V1B2: logFC=+0.99, p=6.46e-58
V-ATPase regulatory	↓ Downregulated	ATP6V1H	logFC=-0.47, p=5.03e-30
TPC channels	↓ Downregulated	TPCN1, TPCN2	TPCN1: -0.32, *TPCN2: -0.42*
TRPML channels	↓ Downregulated	MCOLN1, MCOLN2, MCOLN3	MCOLN1: -0.20, MCOLN2: -1.34, MCOLN3: -0.82
Lysosomal K+ channel	↓ Downregulated	TMEM175	logFC=-0.018*
Lysosomal Cl- channel	↓ Downregulated	CLCN7	logFC=-0.027*

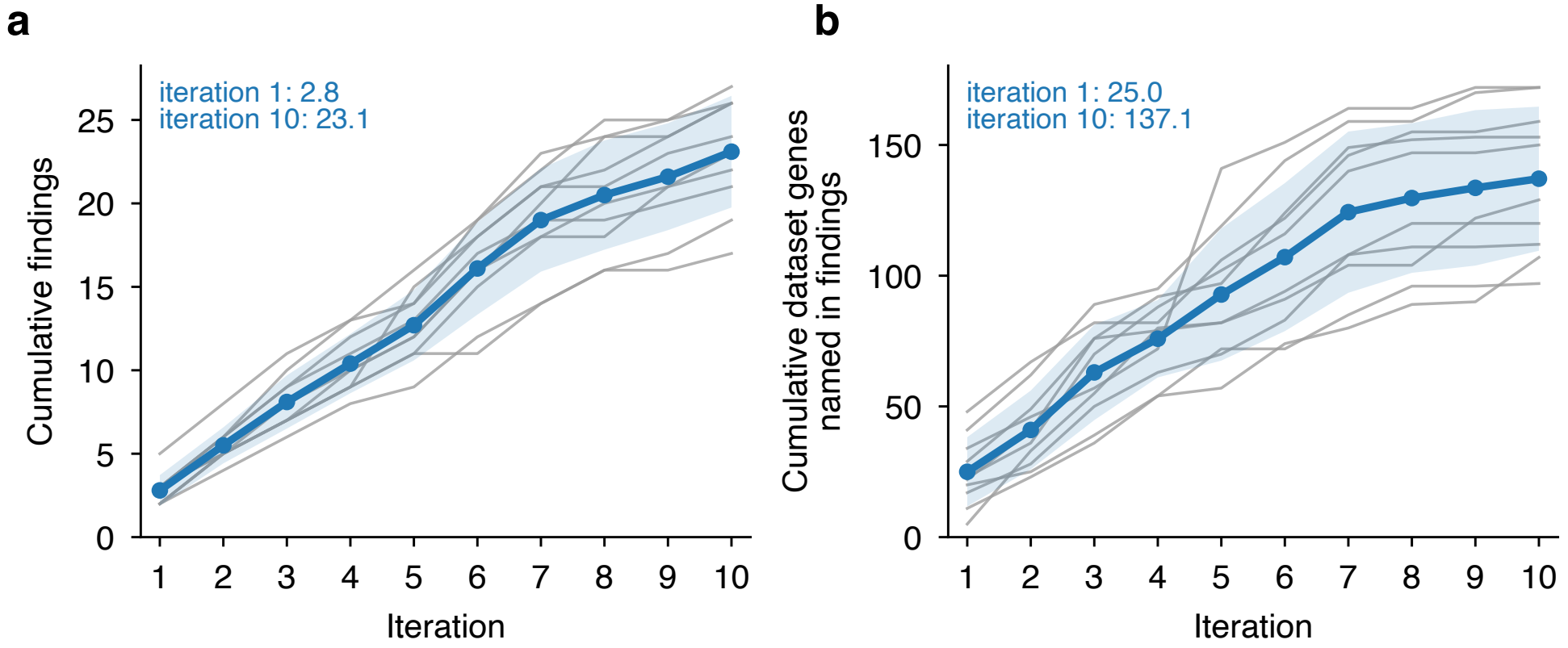
OpenScientist-generated analysis

"The analysis revealed a fundamental discovery: lysosomal acidification is impaired not through V-ATPase deficiency, but through a uniform downregulation of lysosomal ion channels (MCOLN1-3, TMEM175, CLCN7, TPC1/2) that prevents proper counter-ion flux required for proton pump function."

c




Autophagy / mitophagy
Chaperones
ISR / UPR
Lyso ion channels
UPS
V-ATPase V0
V-ATPase V1
V-ATPase assembly
mTOR / Ca / kinases
benchmark
rep-1-of-9
rep-2-of-9
rep-3-of-9
rep-4-of-9
rep-5-of-9
rep-6-of-9
rep-7-of-9
rep-8-of-9
rep-9-of-9
ATG5
ATG7
LAMP1
NPC1
SQSTM1
TFEB
CTSB
LAMP2
MAP1LC3B
PSAP
STX17
BNIP3
TFE3
CTSD
GABARAPL2
PINK1
ULK2
GABARAP
SNAP29
VAMP8
GABARAPL1
BECN1
TOLLIP
ULK1
MAP1LC3A
NBR1
OPTN
MITF
WIPI2
HSP90AA1
FKBP5
HSP90AB1
HSPA1A
HSPA5
HSPA8
HSPB1
DNAJA1
DNAJB6
FKBP4
STIP1
BAG3
CRYAB
DNAJB1
HSPA1B
ATF4
DDIT3
EIF2AK3
PPP1R15A
XBP1
ERN1
CHAC1
PSAT1
TRIB3
ATF3
MCOLN1
TMEM175
CLCN3
CLCN7
CLCN4
MCOLN2
MCOLN3
TPCN1
TPCN2
SLC9A6
CLCN6
UBC
PRKN
VCP
UCHL1
UBB
PSMA1
STUB1
PSMB5
PSMD11
SEL1L
SYVN1
UCHL5
ATP6AP1
ATP6AP2
ATP6V0A1
ATP6V0D1
ATP6V0B
ATP6V0C
ATP6V0E2
ATP6V0E1
ATP6V0A2
ATP6V0A4
ATP6V0D2
ATP6V1H
ATP6V1B2
ATP6V1G1
ATP6V1C1
ATP6V1F
ATP6V1D
ATP6V1A
ATP6V1E1
ATP6V1C2
ATP6V1G2
ATP6V1B1
ATP6V1E2
ATP6V1G3
DMXL2
DMXL1
WDR7
VMA21
CCDC115
TMEM199
ROGDI
RABCONNECTIN
LAMTOR1
RRAGA
RRAGC
MAPT
RHEB
MTOR
NFE2L2
RPTOR
CALM1
FTL
RYR2
CDKSR1
FTH1
CDKSR2
MEF2C
PPP3CA
LRRK2
CAST
SLC7A5
CAMK2A
RBFOX1



Case Study 3, 10 replicate runs (grey); mean ± s.d. (blue)

a

Phase 1. Discovery and hypothesis generation

Discovery Microarray Dataset (GSE47552)
López-Corral et al. (n = 99)

NPC (n=5)	MGUS (n=20)	SMM (n=33)	MM (n=41)

OpenScientist AI Platform

Instructions:
Generate hypotheses using differential gene expression and functional enrichment analysis

Output
Candidate biological hypothesis

Phase 2. Experimental validation design

External validation dataset (GSE6477)
Mayo Clinic (n = 162)

Dataset design

Path A: Authentic validation set with biological signals (preserved data)

Path B: Negative control (randomized data labels with no biological signals)

Phase 3. AI validation run and results

OpenScientist

Instructions:
Validate hypotheses from phase 1 across paths A and B

Output

Path A: Authentic dataset validation
Signal detected and hypothesis confirmed

Path B: Randomized dataset validation
No signal detected in the negative control

Conclusion

Platform correctly distinguishes true biological signal from technical noise

b

OpenScientist-generated table

Hypothesis	Authentic Dataset V	Randomized Dataset R	Conclusion
H1: MGUS → SMM is critical	REJECTED	REJECTED	Normal → MGUS is critical
H2: UPR failure	Supported	Not supported	Partial
H3: Widespread DEGs	Supported	Not supported	Partial
H4: Drug targets	Supported	Not supported	Partial

c

OpenScientist-generated analysis

"Dataset R may have undergone processing that obscured biological signal"

"Technical analysis revealed that Dataset V has a 6.9-fold higher signal-to-noise ratio than Dataset R, likely explaining the differential results."

Run	Run ID	Main finding	Secondary	Statistical considerations	Limitations
Benchmark	623261e1	Impaired acidification via impaired assembly (ATP6V1H)	Loss of compensatory ion channels (Finding 6)		Transcriptomic only analysis, cross-sectional design, selection bias
Replicate 1	790c3cd8	Impaired acidification via impaired assembly (ATP6V1H)	Loss of compensatory ion channels (Finding 3)	Flagged that sample size limited pseudobulk analyses (flagged that some findings do not reach significance with pseudobulking)	Transcriptomic only, cross-sectional design, selection bias
Replicate 2	8bed6957	Impaired acidification via impaired assembly (ATP6V1H)	Loss of compensatory ion channels (Finding 5)	Confirmed that all findings replicate across patients and survive 1000-fold permutation testing	Transcriptomic only; cross-sectional design; small effect size of some findings
Replicate 3	b2b90e80	Impaired acidification via impaired assembly (ATP6V1H)	Loss of compensatory ion channels (Finding 4)	Tested consistency across the 8 donors	Transcriptomic only; cross-sectional design
Replicate 4	618d9649	Impaired acidification via impaired assembly (ATP6V1H)	Loss of compensatory ion channels (Finding 13)	Specific investigation of statistical rigor: consistency across donors, pseudobulking	Transcriptomic only; cross-sectional design
Replicate 5	ba590170	Impaired acidification via impaired assembly (ATP6V1H)		Checked consistency across neuronal subtypes	Transcriptomic only; cross-sectional design; modest effect size
Replicate 6	9e4de575	Impaired acidification via impaired assembly (ATP6V1H)		Specific investigation of statistical rigor: consistency across donors, pseudobulking	Transcriptomic only, cross-sectional design, selection bias
Replicate 7	c19a0a06	Impaired acidification via impaired assembly (ATP6V1H)	Loss of compensatory ion channels (executive summary)	Specific investigation of statistical rigor: consistency across donors, pseudobulking, mixed-effects model accounting for patients, sequencing depth and neuronal subtype identity	Transcriptomic only
Replicate 8	62e0db89	Impaired acidification via impaired assembly (ATP6V1H)		Specific investigation of statistical rigor: mixed-effects model accounting for patients, sequencing depth and neuronal subtype identity	Transcriptomic only; cross-sectional design
Replicate 9	3af7fbd3	Impaired acidification via impaired assembly (ATP6V1H)		Tested consistency across the 8 donors	Transcriptomic only; cross-sectional design

ACCELERATED ARTICLE PREVIEW

Configuration	Network Mode	Run IDs	C1	C2	C3	FDR	Donor-aware
Claude Code + Opus 4.8	online	8efbe47f, 32703f2b, 672938a7, ce0d1c6c, 4effee04, 84f4ac0f, 1abbd1c2, e9ef48c7, a9cda9b8, 8ddde7f1	10/10	10/10	10/10	10/10	10/10
Claude Code + Opus 4.8	air-gapped	9016120e, e3ce8efc, a505cc39, 7638ba4d, f50ebeda, 42b43b4c, 2e7880e0, 937b8a59, 98b3d8fe, da119088	10/10	10/10	10/10	10/10	10/10
omp + Kimi K3	online	7e19b596, fca12230, d3cc57dd, 222863fb, c6f90962, c3d976f3, 73591085, 9dc82985, b71d61f2, 0972151d	10/10	10/10	10/10	10/10	10/10
omp + Kimi K3	air-gapped	fe49c0b4, 801e7329, 28995e5f, d8849ae1, 6848d905, 427622f0, f1f28fbd, 8135d7d8, 3e454378, 49b02891	10/10	10/10	10/10	10/10	10/10
omp + GLM 5.2	online	d349f0b9, 3b4137e1, 3f1268be, bd6cf7db, 13ee5bb3, ca02622e, 8e31ca18, 432129b0, 01bbf49f, c139fa4f	10/10	10/10	10/10	8/10	10/10
omp + GLM 5.2	air-gapped	fb219894, bd105fe1, 99687072, d80d0574, 99e65748, 96d48592, 9821dab9, a6ebde3b, c040e49e, 32960996	10/10	10/10	10/10	9/10	10/10

ACCELERATED SAMPLE PREVIEW

Matched finding	Present in A and B: human adjudicator	Present in A and B: GPT adjudicator	Present in A and B: Claude adjudicator	Consensus relationship between human (A) and OpenScientist (B) analyses
No individual proteins survive FDR	Yes	Yes	Yes	Concordant
CSF T16 broad pathway enrichment	No	No	No	Not identified in B
CSF T16 translation/RNA processing	No	No	No	Not identified in B
CSF T16 nucleotide/nucleoside-binding biology	No	No	No	Not identified in B
CSF T16 endocytosis	No	No	No	Not identified in B
CSF T16 Metascape/MCODE modules	No	No	No	Not identified in B
CSF T24 exploratory pathway enrichment	No	Partial	Partial	Limited overlap. A reports multiple exploratory CSF T24 pathways; B identifies some related inflammatory signal, which GPT and Claude considered partial overlap, while the human adjudicator did not.
Plasma T16 broad metabolic pathway enrichment	No	No	No	Not identified in B. A reports broad metabolic enrichment, whereas B finds a depletion rather than an excess of nominally significant proteins at plasma T16.
Plasma T16 organic/carboxylic acid metabolism	No	No	No	Not identified in B
Plasma T16 amino-acid metabolism	No	No	No	Not identified in B
Plasma T16 endocytosis downregulation	No	No	No	Not identified in B
Plasma T24 proteostasis/autophagy/vesicle-trafficking response	No	Yes	Partial	Both analyses identify a strong plasma T24 signal, with substantial biological overlap in proteostasis, autophagy, and vesicle trafficking. The adjudicators differed in whether A's broader pathway result and B's focused proteostasis result constituted the same finding.
Temporal CSF T16 to plasma T24 pattern	No	Partial	Partial	Partial concordance. A interprets earlier CSF and later plasma pathway changes as a temporally ordered response; B also identifies time-dependent signals but does not support a sequential CSF to plasma process.
CSF-to-plasma clearance / central-to-peripheral efflux	No	No	No	Discordant interpretation. A proposes enhanced CSF-to-plasma clearance as a mechanism; B directly tests cross-compartment protein behavior and finds little evidence for simple CSF-to-plasma redistribution.
CSF mTOR up / Beclin-1 down autophagy-related pattern	No	Partial	Partial	Both analyses contain similar mTOR/Beclin-1 protein observations, but the human adjudicator catalogued these as unique to B. Claude explicitly distinguishes agreement in the protein data from disagreement/uncertainty in the autophagy interpretation.
Plasma T24 STUB1/CHIP-proteostasis mechanism	No	Yes	Yes	Partially concordant. Both analyses identify a plasma T24 protein-processing/proteostasis response, but A emphasizes broad pathway-level changes whereas B emphasizes STUB1/CHIP and related chaperone proteins; the human adjudicator treated these as distinct findings.

ACCELERATED ARTICLE PREVIEW

nature portfolio

Corresponding author(s): Bateman & Reese

Last updated by author(s): Nov 5, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection; all datasets were pre-existing (see Data Availability).
Data analysis	Analyses used OpenScientist (Apache 2.0; https://github.com/openscientist-io/openscientist) with Claude Code (2.0.37 for the featured case studies; 2.1.228 for the supplementary experiments) or omp 17.1.5 agents and the LLMs (and versions) listed in Methods and Supplementary Table 1. OpenScientist wrote and executed its analyses in Python 3.11 (featured case studies: pandas 2.3, numpy 2.4, scipy 1.16 to 1.17, scikit-learn 1.8, statsmodels 0.14.6, matplotlib 3.10) or Python 3.12.13 (supplementary experiments: pandas 3.0.0, numpy 2.3.5, scipy 1.16.3, scikit-learn 1.8.0, statsmodels 0.14.6, scanpy 1.12, anndata 0.12.6, matplotlib 3.10.8, seaborn 0.13.2). Human comparison analyses used SAS 9.4 (Case Study 1), and Metascape v3.5 (metascape.org, Reactome pathways, run before January 2026) with an elastic net model implemented with scikit-learn (Case Study 8). Analysis scripts for the figures and tables are deposited with the case-study data on Zenodo (https://doi.org/10.5281/zenodo.18852838).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

For the featured main text and supplementary case studies, the query, the input data (where possible), and all outputs are available in the Supplementary Material and on Zenodo (https://doi.org/10.5281/zenodo.18852838). SEABIRD and the suvorexant study are unpublished. Data can be obtained with approval under data use agreement from the corresponding authors (Randall Bateman and Brendan Lucey). All other datasets are public (CELLxGENE; GEO accessions GSE6477 and GSE47552; Harvard Dataverse doi:10.7910/DVN/8S7JVS; NF Data Portal, RRID:SCR_021683). Source Data are provided for Figures 2 to 4 and Extended Data Figs. 1 and 2.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Sex was available as a demographic variable in the SEABIRD and Biomarkers of Aging datasets and was used as a covariate in the Case Study 2 baseline model, but no sex-based analyses were performed because the study evaluated an AI analysis platform rather than sex-specific hypotheses.
Reporting on race, ethnicity, or other socially relevant groupings	Race was available as a demographic variable in SEABIRD, a community cohort recruited for diversity, but was not used to test hypotheses about group differences.
Population characteristics	Datasets comprised 325 SEABIRD participants with plasma biomarkers and amyloid PET (Case Study 1), 500 Biomarkers of Aging participants aged 18–99 (Case Study 2), post-mortem brain tissue from 8 donors (Case Study 3), 99 and 162 plasma cell disorder patients (Case Study 4), and 25 cognitively unimpaired suvorexant trial participants (Case Study 5); full characteristics are reported in the original publications.
Recruitment	No participants were recruited for this study; all data came from existing cohorts, trials and public datasets whose recruitment is described in the original publications.
Ethics oversight	SEABIRD and the suvorexant study were approved by the Washington University in St. Louis IRB (SEABIRD, IRB 201902081, NCT03899844; suvorexant, IRB 201908156, NCT03077620) with written informed consent from all participants.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No sample-size calculation was performed; dataset sizes were determined by existing cohorts, all development-period runs are reported (Fig. 2, Supplementary Table 1), and repeated-run experiments used 10 runs per configuration (15 per arm for the skills comparison) to characterize variability.
Data exclusions	Failed runs and runs without data were excluded from case-study selection (Fig. 2), and dataset-level exclusions and data handling (duplicate SEABIRD records removed, Case Study 2 participant handling by run, one preregistered Case Study 5 dataset not adjudicated owing to time constraints) are described in Methods.
Replication	Case Study 1 was run 11 times, Case Study 3 was replicated more than 100 times across models, agents, network modes and ablations with consistent central conclusions, Case Study 2 was evaluated on a hidden held-out subset, and Case Study 4 hypotheses were tested in an independent external cohort.
Randomization	Case Study 2 used a fixed-seed, mortality-stratified train/evaluation split, Case Study 4 used disease-label randomization as a negative control, and Case Study 5 participants were randomized to drug or placebo in the original trial; other analyses involved no group allocation.

Blinding	In Case Study 5, de-identified human and OpenScientist reports were adjudicated blind by an independent domain expert; OpenScientist was blinded to the hidden evaluation set in Case Study 2 and to label randomization in Case Study 4, and expert evaluations in Case Studies 1–3 were not blinded.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	This study did not register a new clinical trial. It used existing data from two clinical trials that have not yet been published: SEABIRD (NCT03899844) and the suvorexant study (NCT03077620). The Case Study 5 analysis plan was preregistered at AsPredicted (#298545) before AI analysis.
Study protocol	Protocols for SEABIRD and the suvorexant study are available from the corresponding authors on request. The Case Study 5 analysis plan is publicly available at AsPredicted (#298545).
Data collection	This study did not collect new clinical data; it performed secondary analyses of existing data from two unpublished clinical trials. SEABIRD participants were recruited from the St. Louis community and primary care settings and underwent blood collection, cognitive assessment and amyloid PET/MRI at Washington University in St. Louis. Suvorexant study participants underwent serial CSF and plasma collection at Washington University in St. Louis.
Outcomes	No new clinical outcomes were defined. The outcomes of this study were agreement between OpenScientist and expert analyses: ROC AUC for amyloid PET status (Case Study 1) and adjudicated finding-level concordance with the preregistered analysis plan (Case Study 5).

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*