Research Square

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

AtlasAI: Multi-Agent Reasoning for Knowledge Discovery in the Human Protein Atlas

Emre Green
Science for Life Laboratory, KTH Royal Institute of Technology
Kalle von Feilitzen
Science for Life Laboratory, KTH Royal Institute of Technology
Fredric Johansson
Science for Life Laboratory, KTH Royal Institute of Technology
Mattias Forsberg
Science for Life Laboratory, KTH Royal Institute of Technology
Zeynep Sumer
King's College London
Xiya Song
Science for Life Laboratory, KTH Royal Institute of Technology
Xinmeng Liao
Science for Life Laboratory, KTH Royal Institute of Technology
Mengzhen Li
Science for Life Laboratory, KTH Royal Institute of Technology
Özlem Altay
Science for Life Laboratory, KTH Royal Institute of Technology
Hong Yang
Science for Life Laboratory, KTH Royal Institute of Technology
Tonghua Zhang
Science for Life Laboratory, KTH Royal Institute of Technology
Xiangtai Kong
Science for Life Laboratory, KTH Royal Institute of Technology
Xuexin Li
Karolinska Institute
Adil Mardinoglu
Science for Life Laboratory, KTH Royal Institute of Technology
Mathias Uhlén
Science for Life Laboratory, KTH Royal Institute of Technology
Cheng Zhang
cheng.5.zhang@kcl.ac.uk

King's College London

Article

Keywords:

Posted Date: April 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9452188/v1

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

Additional Declarations: There is **NO** Competing Interest.

Abstract

The Human Protein Atlas provides expression and localization data for 20,000 genes, yet extracting insights requires complex queries. We present AtlasAI, a multi-agent system deployed as a sidebar within the atlas enabling natural language querying and analysis. Specialized agents for search, gene investigation, refinement, and interpretation produce auditable traces with semantic validation and error recovery. Across a 236-query benchmark spanning six biological constraints, AtlasAI achieved complete task resolution, whereas frontier models showed lower performance at higher cost. We introduce the Autonomous Scientific Orchestrator, which accepts a research objective, independently coordinates agents through tool selection and code-level validation, and eliminates hallucinated data by constraining outputs to externally computed artifacts. In evaluation, it identified 486 genes, retrieved 1,006 measurements, and generated 12 publication-ready visualizations without human intervention.

Introduction

The Human Protein Atlas (HPA) represents a comprehensive mapping of the human proteome across tissues, organs, cells, and subcellular compartments *(1–3)*. The atlas encompasses expression data from immunohistochemistry across 44 normal human tissues, RNA sequencing from 37 organs, single-cell transcriptomics spanning hundreds of cell types, protein localization patterns in 35 subcellular structures *(4)*, expression profiles across 64 cell lines, and disease associations including 17 major cancer types *(2, 5)*. The Blood Atlas provides blood-cell expression profiles *(6)*, the Secretome Atlas distinguishes proteins secreted to blood from those retained locally in tissues *(7)*, and the Pathology Atlas links gene expression to patient survival outcomes across cancers *(2)*. Together, these resources contain over 26 million immunohistochemistry images, expression values for 20,162 genes, and structured annotations enabling sophisticated biological queries.

This wealth of information has enabled discoveries ranging from tissue-specific protein signatures *(8)* to principles of subcellular organization *(9)* to cancer prognostic markers *(10)*. Yet the very richness that makes the Human Protein Atlas valuable also makes it challenging to navigate. The database organizes filterable attributes across multiple categories (Tissue, Brain, Cancer, Single Cell, Subcellular, Cell Line, Gene Info, Structure, Interaction), each containing hierarchical fields, classes, and subclasses. Constructing a query like "membrane receptors enriched in brain but absent from blood" requires correct filter syntax across several categories. AtlasAI eliminates this barrier by allowing users to ask questions in natural language through an integrated sidebar deployed directly on the Human Protein Atlas website.

Large language models (LLMs) have demonstrated remarkable capabilities in natural language understanding *(11–13)*. However, applying monolithic language models to scientific knowledge navigation presents fundamental challenges. First, language models are prone to hallucination, generating plausible but factually incorrect outputs *(14, 15)*, a failure mode particularly dangerous in scientific contexts. Second, the reasoning process occurs through billions of parameter interactions in a single forward pass, rendering it effectively opaque *(16)*. Researchers cannot inspect intermediate

reasoning, identify where errors occurred, or verify that correct answers arose through sound logic rather than pattern matching. This opacity conflicts fundamentally with scientific norms of transparency and reproducibility *(17, 18)*.

Agentic AI architectures offer a fundamentally different approach *(19–21)*. Rather than processing queries through a single model invocation, agentic systems decompose tasks into discrete steps executed by specialized components, each producing inspectable intermediate outputs. This decomposition enables explicit reasoning traces, tool use with verifiable outcomes, and iterative refinement based on intermediate results *(22–24)*. Recent work has demonstrated the power of multi-agent approaches for code generation *(25)*, web navigation *(26, 27)*, and scientific reasoning *(28, 29)*, consistently showing that architectures with multiple specialized agents outperform monolithic alternatives while providing interpretable decision traces.

Here, we present a multi-agent system for autonomous navigation of the Human Protein Atlas. Our architecture employs specialized agents for distinct query types, each operating through explicit reasoning steps that generate complete audit trails. Beyond single-query interactions, the Autonomous Scientific Orchestrator (ASO) chains these agents with analytical and visualization tools to execute multi-step research objectives from a single natural language prompt, persisting every intermediate output in a structured artifact store that serves as both the communication backbone between tools and a complete audit record.

Results

Multi-agent architecture for transparent biological reasoning

The fundamental limitation of monolithic language models for scientific applications is the opacity of their reasoning process *(16, 30)*. When a user submits a query, the model produces an output through a single forward pass involving billions of parameters, with no visibility into intermediate computations or decision points. If the output is incorrect, there is no mechanism to identify where reasoning failed. Our multi-agent architecture addresses this through explicit decomposition of biological queries into specialized processing pathways, each generating auditable intermediate outputs. The contrast between monolithic and multi-agent approaches is illustrated in Fig. 1A: under direct prompting conditions, monolithic models produced outputs through opaque forward passes with 0 to 20% accuracy, whereas our architecture generated inspectable intermediate outputs at each reasoning step, achieving complete task resolution at 56 to 240 times lower computational cost (Fig. 2A, Table S3).

The architecture comprises five agents designed for fundamentally different query types (Fig. 1B). The Deep Research agent processes complex searches requiring construction of multi-constraint database queries, such as "find kinases enriched in brain tissue but not secreted to blood." The Investigator agent handles queries requiring extraction of specific information from within gene pages and interactive

elements, such as "is EGFR prognostic in lung cancer?" The Inclusion Check agent enables iterative exploration by verifying whether specific information appears in previous search results, supporting workflows where researchers progressively refine their queries. The Dictionary Expert agent handles educational queries about tissue histology, cancer pathology, and cell biology, mapping natural language questions to the Human Protein Atlas structured dictionary of annotated microscopy images across all sub-atlas pages. The Autonomous Scientific Orchestrator (ASO) coordinates these agents with analytical and visualization tools to execute fully autonomous multi-step research workflows from a single natural language prompt (Fig. 1C). Agent selection occurs automatically based on query analysis, with every routing decision logged for transparency. The complete system is deployed as an integrated sidebar on the Human Protein Atlas website, streaming all reasoning steps to the user interface in real time (Fig. S1).

Deep research agent: five-stage query pipeline

The Deep Research agent processes complex biological queries through five explicit stages (Fig. 1A), each generating logged intermediate outputs that enable complete reconstruction of the reasoning pathway. This multi-stage architecture reflects insights from recent work on chain-of-thought reasoning *(22)* and tool-augmented language models *(23)*, extended to the domain of structured biological knowledge navigation.

In the first stage, Goal Parsing, the agent decomposes natural language queries into structured components: subject (what class of genes or proteins is sought), location (where expression is required), qualifier (how expression should be characterized), and exclusion (where expression should be absent). The parser also generates synonyms and qualifier variants to improve downstream trail scoring. In Constraint Planning, the agent determines how parsed components map to database operations, reasoning about component interactions. A query for "FDA approved kinases" requires two independent protein class constraints (Enzymes;Kinases AND FDA approved drug targets), while a query for "proteins secreted in brain" may require only a single secretome constraint (Fig. 3C). The constraint plan is logged as an ordered list of AND/NOT operations.

In Trail Navigation, the system scores all 842 available class-level paths using a hybrid scoring function that combines token overlap, synonym matching, and prefix matching against the full hierarchical schema, presenting candidates per component to the language model for selection. This approach ensures the language model always sees the complete candidate space rather than relying on automated pre-filtering that may discard relevant options. In URL Construction, selected paths are assembled into the executable compound search URL with correct field encodings and operator placement. In Semantic Validation, the constructed query is compared against the original intent through an independent language model invocation that produces a structured verdict (PASS, AUGMENT, or REJECT) with per-component checking (Fig. 3A). On REJECT, the agent autonomously re-plans and re-executes with corrected parameters. The validation stage accounts for approximately one third of total

computation, reflecting the architecture emphasis on ensuring query correctness before returning results.

Investigator agent: evidence-grounded gene queries

The Investigator agent handles gene-specific queries through autonomous page exploration and evidence-grounded synthesis (Fig. 3B). The agent resolves gene symbols to Ensembl identifiers, selects relevant HPA pages (Tissue, Brain, Cancer, Single Cell, Subcellular), and extracts quantitative data from interactive charts, tables, and bubble plots through direct HTML parsing. The extraction pipeline handles bar charts with grouped categories, bubble plots encoding tissue by cell type expression matrices, and paginated data tables, automatically detecting units (normalized transcripts per million (nTPM), normalized counts per million (nCPM), fragments per kilobase of transcript per million mapped reads (FPKM)) from tooltip annotations. An entity-centric merge stage deduplicates identical charts across page sections, strips zero-expression entries, and groups data by legend labels, reducing the information presented to the language model while preserving all non-trivial expression values.

The language model synthesizes answers exclusively from extracted data, architecturally preventing fabricated statistics. A validation stage confirms that all cited values are grounded in the extracted evidence. This separation of "what to look for" (where language model capabilities are valuable) from "what was found" (which must be extracted from authoritative sources) provides stronger guarantees than prompting strategies alone *(31, 32)*. On a curated benchmark of 17 queries spanning simple searches, compound filters, exclusion operations, gene investigations, and histology education (Table S1), AtlasAI achieved complete task resolution under the evaluation criteria, with the validation stage recovering all four queries requiring intervention.

Benchmark performance

We evaluated AtlasAI on a 236-query benchmark spanning five complexity tiers defined by the number of simultaneous constraints (Table S4). Queries ranged from single-constraint searches (for example, "prognostic markers in melanoma") to six-constraint queries combining protein class intersections, tissue enrichments, brain region specificity, cell type lineages, subcellular locations, and secretome exclusions (for example, "FDA drug targets enriched in brain, not detected in liver, kidney, or heart, and not secreted to blood"). The benchmark was heavily weighted toward complex queries: 130 queries required three simultaneous constraints, 32 required four, and 37 required five or more, with 77% of all queries including at least one exclusion constraint (NOT operator).

Across all 236 queries, AtlasAI achieved complete task resolution across all benchmark queries under predefined evaluation criteria with a mean completion time of 13.9 seconds. The first-attempt success rate was 95.8%, with the semantic validation mechanism successfully recovering all queries requiring intervention. Query types spanned protein classes (enzymes, kinases, transcription factors,

transporters), tissue expression categories (RNA enrichment across 37 organs), brain region specificity (17 regions), single-cell type enrichment (over 80 cell types), immune cell lineage (T-cells, B-cells, NK-cells), subcellular locations (35 compartments), secretome annotations, prognostic cancer markers (The Cancer Genome Atlas (TCGA) formal nomenclature), and compound protein class intersections such as "cancer-related kinases" requiring two protein_class filters joined by AND. Every constructed URL was verified against live HPA search results, with row counts confirmed to match (Table S4).

Comparison with frontier language models

To evaluate whether specialized multi-agent architecture provides advantages over direct prompting of frontier language models, we benchmarked AtlasAI against six model configurations: Claude Opus 4.6 with extended thinking, Gemini 3.1 Pro with thinking, GPT-5.4 at high, medium, and no reasoning effort, and GPT-4.1 (Fig. 2A, Table S3). Each model was given identical queries with access to web browsing capabilities and instructed to answer biological questions using the Human Protein Atlas as their data source.

AtlasAI achieved complete task resolution across all 30 benchmark queries under predefined criteria. In contrast, under direct prompting conditions, five of six model configurations failed to produce any correct answers. Claude Opus 4.6 and GPT-5.4 (at all three reasoning levels) attempted to access the Human Protein Atlas through web browsing, tried multiple strategies including programmatic API access and bulk data downloads, but ultimately could not navigate the hierarchical search system and explicitly declined to provide answers rather than risk fabrication. GPT-4.1, lacking reasoning capabilities, produced no correct answers. Gemini 3.1 Pro achieved 20% task success under the same evaluation criteria (6 of 30 correct within a 10% tolerance threshold), but unlike the models that refused, Gemini provided answers for all 30 queries with high stated confidence, hallucinating gene counts for four queries where the Human Protein Atlas returns zero results and fabricating specific gene names (Fig. 2B). The failure modes revealed two distinct limitations: models either recognized they could not reliably execute the task and refused, or approximated from parametric knowledge rather than navigating the actual database schema. Neither approach produced usable results. The Atlas Cost Index, measuring API cost relative to AtlasAI, ranged from 56-fold (GPT-5.4 without reasoning) to 240-fold (Claude Opus 4.6), demonstrating that multi-agent architecture achieves superior accuracy at dramatically lower cost *(33, 34)*.

Time efficiency

Timing measurements comparing AtlasAI against manual navigation by experienced HPA researchers revealed acceleration that increases with query complexity (Fig. 2C, Table S2). Simple single-constraint searches showed 3.1-fold speedup (37.5 seconds manual versus 12.1 seconds autonomous). Compound multi-constraint searches showed 9.0-fold speedup (131.3 seconds versus 14.5 seconds), reflecting the substantial cognitive and mechanical overhead of constructing multi-constraint queries

through hierarchical menus. Exclusion queries showed 9.8-fold speedup (178.0 seconds versus 18.1 seconds), as NOT operations require particularly careful attention to filter construction. Gene investigations showed 5.5-fold acceleration (52.7 seconds versus 9.5 seconds). These efficiency gains were achieved while increasing rather than decreasing transparency: each query generated between 8 and 28 logged reasoning steps, providing complete auditability of the navigation process *(35)*.

Autonomous scientific orchestrator

The Autonomous Scientific Orchestrator (ASO) extends single-query navigation into fully autonomous multi-step research workflows (Fig. 4A). Unlike prior agent frameworks, ASO externalizes all intermediate state into a persistent artifact store, preventing the language model from generating or modifying data directly. A natural language research objective enters the LLM orchestrator, which iteratively selects tools from a 14-tool registry spanning five categories: agent summoning (Deep Research, Investigator, Dictionary Expert, Inclusion Check, Inspect), data cleaning and measurement (clean, compare, measure with scout-and-direct pipeline), analysis (rank, delta, merge, scatter, matrix, concat), visualization (15 chart types), and workflow completion. The orchestrator operates through a think-act-store-observe loop: at each iteration, it reads the current workspace state, reasons about the next action, calls tools, persists results as UUID-addressed artifacts, and observes the updated workspace snapshot before planning the next step.

Three architectural decisions distinguish ASO from conventional language model pipelines. First, all inter-tool communication routes through a persistent artifact store rather than passing data through the language model context, eliminating a class of hallucination errors while reducing token consumption *(36)*. The orchestrator sees only artifact previews containing column names, row counts, and sample values. Second, code-level guardrails enforce artifact kind constraints at every tool invocation, rejecting invalid operations (for example, passing ranked data to a scatter tool) with machine-readable recovery hints that enable autonomous self-correction. Third, the scout-and-direct measurement pipeline uses a single language model call to discover page structure for the first gene, then measures all subsequent genes through direct HTTP page fetching and parsing with zero additional language model calls, achieving approximately 20-fold cost reduction for measurement-intensive workflows *(37)*.

We evaluated ASO on six research queries of increasing complexity (Table 1). Across all runs, the system achieved 100% completion, autonomously measuring 486 unique genes across 1,006 data points and producing 12 interactive charts spanning five visualization types. Easy queries (direct multi-gene comparisons) completed in 3 to 4 steps, while complex comparative analyses required 7 to 10 steps with up to 534 individual measurements per run. The system demonstrated robust error recovery: in the most complex benchmark (114 enzymes, 3 chart types), an initial scatter analysis call was correctly rejected by the kind validation guardrail, and the orchestrator autonomously reformulated the request using raw measurement identifiers on the subsequent step (Fig. 4B). The complete workspace for each

run, including all structured datasets, generated visualizations (Fig. 4C), and the full execution log, is downloadable as a portable archive.

Discussion

The application of artificial intelligence to scientific knowledge discovery presents a fundamental tension: AI systems can dramatically accelerate research, yet the opacity of neural network reasoning conflicts with scientific norms of transparency and reproducibility *(31, 32, 38)*. Our work demonstrates that multi-agent architectures resolve this tension.

The direct comparison with frontier language models is instructive. Six model configurations spanning four model families and three reasoning effort levels all failed to reliably navigate the Human Protein Atlas, despite having web browsing access and extended reasoning capabilities. The dominant failure mode was architectural: monolithic models cannot reliably translate natural language into precise hierarchical filter syntax, instead approximating from parametric knowledge. The 56 to 240-fold cost premium these models carry makes the failure even more striking: architecture, not scale, determines success at structured knowledge navigation *(39, 40)*.

The semantic validation mechanism provides capabilities impossible in monolithic architecture. Comparing constructed queries against original intent and triggering retry cycles on mismatch requires explicit intermediate representations unavailable in single-model approaches. The Investigator agent's mandatory evidence grounding, which architecturally prevents fabricated statistics, relies on separating the "what to look for" decision from the "what was found" evidence. These architectural constraints provide stronger guarantees than prompting strategies or fine-tuning objectives, which can always be circumvented by sufficiently adversarial inputs *(41, 42)*.

The Autonomous Scientific Orchestrator demonstrates that multi-agent architectures can scale from single-query tools to autonomous research pipelines without sacrificing transparency. The key architectural decision of routing all inter-tool communication through a persistent artifact store simultaneously eliminates hallucinated data generation and makes every intermediate result inspectable. Code-level guardrails that enforce artifact kind constraints proved more reliable than prompt-based instructions, as demonstrated by the automatic error recovery in the most complex benchmark. The scout-and-direct measurement pipeline enables workflows spanning hundreds of genes at practical cost *(43, 44)*.

The implications extend beyond the Human Protein Atlas. The multi-agent architecture we describe, comprising goal decomposition, hierarchical navigation with hybrid scoring, semantic validation with retry logic, and evidence-grounded synthesis, is applicable to any structured knowledge base with hierarchical filters. Databases such as UniProt *(45)*, Ensembl *(46)*, Gene Ontology *(47)*, and many others present similar navigation challenges. The core principles of explicit reasoning traces, intermediate validation, and evidence grounding are database-agnostic and represent general solutions to the transparency challenge in AI-augmented scientific discovery *(48, 49)*.

Several limitations warrant discussion. Natural language permits inherently ambiguous formulations; while our validation mechanism catches many misinterpretations, some queries genuinely require clarification. Query outputs are sensitive to natural language phrasing, reflecting the system's faithful translation of user intent into the closest available database filter, consistent with how an experienced researcher would interpret the same phrasing when constructing a manual search. The hierarchical navigation approach assumes structured filter systems; unstructured databases would require adaptations. The evidence grounding constraint in the Investigator agent prevents synthesis across multiple sources. ASO demonstrates autonomous multi-step workflows with 100% completion on the current benchmark. As artificial intelligence becomes increasingly integrated into research workflows, the challenge is ensuring outputs are produced through verifiable, reproducible processes. Multi-agent architectures provide a paradigm for AI-augmented science that aligns with rather than undermines scientific methodology *(50, 51)*.

Online Methods

System Implementation and Deployment

AtlasAI is implemented as a Node.js Express server with a React frontend, deployed as an integrated sidebar directly on the Human Protein Atlas website (proteinatlas.org). The sidebar is the first interactive element researchers encounter when visiting any atlas page, providing immediate access to natural language navigation without leaving the HPA interface. All agent reasoning streams to the client via Server-Sent Events, with each reasoning step persisted to a MySQL database as structured JSON for real-time deterministic UI reconstruction. The backend communicates through OpenAI-compatible API endpoints, enabling support for multiple model providers. The Deep Research agent and ASO orchestrator use GPT-4.1 mini as their backbone language model. Conversation history is compacted differentially, preserving user turns while summarizing assistant reasoning to manage context length across extended sessions.

Query Router and Agent Selection

Incoming queries are analyzed to determine the appropriate agent. The router examines whether the query requests a specific piece of information found within a gene’s pages or interactive elements (routing to Investigator), asks about the presence of a gene or protein in previous search results (routing to Inclusion Check), requests histology, pathology definitions, or tissue structure information (routing to Dictionary Expert), initiates a multi-step research workflow requiring autonomous tool orchestration (routing to ASO), or requires database search with extensive filters (routing to Deep Research). The routing decision is logged and streamed to the UI, enabling users to verify that their query was interpreted correctly before processing begins.

Deep Research Agent

The Deep Research agent processes queries through five stages. In Goal Parsing, natural language queries are decomposed into structured JSON components (subject, location, qualifier, exclusion) with explicit field validation. The parser also generates synonyms and qualifier variants for downstream trail scoring. In Constraint Planning, the agent maps parsed components to database operations, determining AND/NOT operator assignments and reasoning about component interactions. Compound subjects (for example, “FDA approved kinases”) are decomposed into separate protein class items requiring two independent filters joined by AND.

In Trail Navigation, the system evaluates all 842 available class-level paths in the HPA schema using a hybrid scoring function. For each query component, the scorer computes token overlap between the search intent and every candidate path, augmented by synonym matching and prefix matching (shared 4+ character prefixes). Up to 120 candidate paths are presented to the language model per component (capped at 25 per category to ensure diversity), and the model selects the most semantically appropriate path. Trail navigation runs in parallel across all query components.

In URL Construction, selected paths are assembled into executable compound search URLs with correct field encodings, subclass notation (semicolon-separated), and operator placement (+AND+ or +NOT+). In Semantic Validation, the constructed URL is compared against the original query intent through an independent language model invocation that produces a structured verdict (PASS, AUGMENT, or REJECT) with per-component checking. The validator checks subject match, location match, qualifier match, and exclusion match independently. On REJECT, the agent re-plans and re-executes with corrected parameters, up to a maximum of four retry cycles.

Investigator Agent

The Investigator agent resolves gene symbols to Ensembl identifiers and selects relevant HPA pages (Tissue, Brain, Cancer, Single Cell, Subcellular, Cell Line, Blood, Interaction, Structure, Pathology) based on query intent. For each selected page, the agent fetches the full HTML and extracts quantitative data through direct DOM parsing of interactive charts (bar charts, bubble plots), data tables, and tooltip annotations. Units (nTPM, nCPM, FPKM, pTPM) are detected from chart axis labels and tooltip text.

An entity-centric merge stage deduplicates identical charts that appear in multiple page sections, strips zero-expression entries, extracts legend group labels from grouped bar charts, and organizes data by entity (tissue, cell type, or subcellular compartment). The merged data is presented to the language model for synthesis, with a mandatory evidence-grounding constraint: the model may only cite values present in the extracted data. A validation stage confirms all cited values against the extraction output.

Dictionary Expert Agent

The Dictionary Expert agent maps natural language questions about histology, pathology, and cell biology to the HPA structured dictionary across nine atlas pages (Tissue Atlas, Brain Atlas, Single Cell, Cancer, Blood Atlas, Cell Line, Subcellular, Structure, Interaction). The agent uses semantic matching to

identify relevant dictionary entries and returns annotated microscopy images with structured descriptions from the HPA knowledge base.

Inclusion Check Agent

The Inclusion Check agent enables iterative exploration by verifying whether specific genes, proteins, or data points appear in results from previous queries within the same session. This supports progressive refinement workflows where researchers narrow their search based on initial findings (for example, first searching for liver-enriched kinases, then checking whether a specific gene of interest appears in that result set). The agent operates on the conversation history and previously returned search results without requiring new database queries.

Semantic Validation Mechanism

The Deep Research validation stage performs an independent language model invocation that receives the original query, the constructed URL, and a human-readable interpretation of each URL component. The validator produces a structured JSON verdict containing: (1) a per-component check (subject, location, qualifier, exclusion) with boolean ok/fail status and reasoning, (2) an overall verdict of PASS, AUGMENT, or REJECT, and (3) corrective guidance when applicable. PASS indicates all components match the original intent. AUGMENT indicates the query is valid but could be expanded (for example, adding an additional tissue). REJECT indicates a semantic mismatch requiring re-planning. On REJECT, the full constraint planning and trail navigation stages are re-executed with the validator's feedback incorporated, up to four retry cycles. The validation stage accounts for approximately one third of total token consumption per query, reflecting the architecture's emphasis on correctness verification.

Autonomous Scientific Orchestrator (ASO)

The ASO orchestrator manages a 14-tool registry organized into five categories: agent summoning (deep_research_hpa, investigator_hpa, inspect), data cleaning and measurement (clean_topx, clean_diff, measure), analysis (analyze_rank, analyze_delta, analyze_merge, analyze_scatter, analyze_matrix, analyze_concat), visualization (chart with 15 chart types), and workflow completion (finish). The orchestrator operates through iterative think-act-store-observe cycles using native function calling.

All inter-tool communication routes through a persistent artifact store. Each tool call produces a UUID-addressed JSON artifact persisted to the workspace filesystem. The orchestrator receives only artifact previews (column names, row counts, sample values), never raw data. Code-level guardrails validate artifact kinds at every tool invocation: for example, analyze_scatter accepts only measurement-kind artifacts, rejecting ranked or merged outputs with machine-readable recovery hints that include the expected kind and available alternatives. This enables autonomous self-correction without external guidance.

The scout-and-direct measurement pipeline optimizes cost for expression measurement workflows. For the first gene in a batch, the system dispatches an Investigator agent call (the "scout") to discover the

page structure, column positions, and value extraction pattern. All subsequent genes in the batch are measured through direct HTTP page fetching and HTML parsing using the discovered pattern, requiring zero additional language model calls. Pages are cached per gene, so measuring multiple tissues for the same gene set incurs only one set of page fetches.

Chart Rendering Pipeline

The ASO chart tool accepts a dataset artifact UUID, chart type, axis mappings, and labels. Chart specifications are passed to a Python rendering script that generates publication-quality PNG images using Matplotlib. The renderer supports 15 chart types: bar, lollipop, diverging bar, dot plot, scatter, bubble, line, grouped bar, stacked bar, box, heatmap, radar, waterfall, volcano, and ridge. Each rendered chart is persisted as a numbered artifact (chart_1.png, chart_2.png, etc.) in the workspace directory. The chart tool validates that the source dataset contains the required columns (x, y, and group/series where applicable) before rendering.

236-Query Deep Research Benchmark

The 236-query benchmark was constructed to span five complexity tiers defined by the number of simultaneous constraints (1 to 6 axes). Queries cover protein classes, tissue expression (RNA enrichment across 37 organs), brain region specificity (17 regions), single-cell type enrichment (over 80 cell types), immune cell lineage (T-cells, B-cells, NK-cells), subcellular locations (35 compartments), secretome annotations, prognostic cancer markers (using TCGA formal nomenclature), and compound protein class intersections. The benchmark deliberately emphasizes complex multi-axis queries: 130 queries require three constraints, 32 require four, and 37 require five to six.

Ground truth was established by verifying each constructed URL against live HPA search results. For each query, the system-generated URL was fetched and the returned row count was compared against the system's reported count. A query was scored as correct if the URL was semantically valid (matching all stated constraints) and the row count matched. The complete benchmark with per-query results is provided in Table S4.

17-Query Curated Benchmark

The curated benchmark comprised 17 queries constructed through consultation with HPA researchers and review of representative user question types. Queries spanned five categories: simple single-constraint searches (n=3), compound multi-constraint searches (n=4), exclusion queries testing NOT operations (n=4), gene-specific investigations (n=4), and histology education queries (n=2). Ground truth was established through manual navigation by three researchers with extensive HPA experience. For search queries, ground truth consisted of the correct URL. For investigation queries, ground truth consisted of specific data values with their page locations. Accuracy was scored as correct if system outputs matched ground truth or returned accurate values with valid citations.

Time Efficiency Measurements

Timing measurements were performed with three expert users familiar with HPA structure, measuring the interval from query formulation to result inspection. For manual navigation, this encompassed the time required to identify appropriate filter categories, navigate the hierarchical menus, construct the query, and verify results. For AtlasAI queries, this encompassed the interval from submitting natural language input to receiving and reviewing results. Each query type was measured across 5 trials per user, and median times are reported.

ASO Benchmark

The ASO benchmark comprised six research queries of increasing complexity (2 easy, 3 medium, 1 hard). Easy queries involved direct multi-gene expression comparisons across specified tissues, requiring measurement and a single chart. Medium queries required autonomous gene list discovery via HPA search followed by cross-tissue measurement, analysis, and chart generation. The hard query combined two independent enrichment searches, quantitative measurement of 114 genes across two tissues, and generation of four chart types including a scatter comparison that triggered guardrail-mediated error recovery. Completion was scored as pass if the system produced all requested visualizations with correct data, and as pass with self-correction (Pass*) when a code-level guardrail rejection triggered successful reformulation.

Benchmark Protocol for Frontier Language Model Comparison

All 30 queries were presented to each model in separate independent sessions. The prompt identified the Human Protein Atlas (proteinatlas.org) as the authoritative data source and instructed each model to provide the complete list of matching genes, the count, and a confidence level for each query. No HPA schema information, URL syntax, or worked examples were provided. All models had web browsing enabled at the time of evaluation. Baseline models were evaluated without access to the HPA schema, reflecting standard usage without task-specific system integration.

Six configurations were tested: Claude Opus 4.6 with extended thinking (Anthropic, $5.00/$25.00 per 1M input/output tokens), Gemini 3.1 Pro with thinking (Google, $2.00/$12.00), GPT-5.4 at high reasoning effort ($2.50/$15.00), GPT-5.4 at medium reasoning effort ($2.50/$15.00), GPT-5.4 with no reasoning ($2.50/$15.00), and GPT-4.1 with no reasoning (OpenAI, $2.00/$8.00). AtlasAI uses GPT-4.1 mini ($0.40/$1.60) as its backbone model. All pricing as of March 2026.

No retries were permitted. Each model received one attempt per query. Models were free to use web browsing, search, and any other tools available in their standard configuration. No time limit was imposed; models were allowed to run until they produced a final answer or explicitly declined. A model was scored as “Refused” when it explicitly stated that it could not reliably complete the task. Correctness was scored within a 10% tolerance of the ground truth count. Atlas Cost Index (ACI) was computed as the estimated API cost per query for each model divided by the AtlasAI cost per query.

Declarations

Data Availability

The complete 236-query benchmark dataset, including per-query URLs, verified row counts, and execution times, is provided in Table S4. Individual execution logs and ASO workspace archives are available from the corresponding authors upon reasonable request. All benchmark queries, evaluation results, and execution logs are provided in the supplementary materials.

Code Availability

The AtlasAI system is accessible at proteinatlas.org. Source code is available from the corresponding authors upon reasonable request.

Acknowledgments

Funding

Knut and Alice Wallenberg Foundation, Sweden, under project number 72110

Author contributions

E.G. designed the multi-agent architecture, developed the software, built the Autonomous Scientific Orchestrator, designed and executed the benchmarks, and wrote the manuscript. F.J. deployed the system on the Human Protein Atlas website. K.v.F. and M.F. provided feedback on system design and testing. S.Z.S., X.S., X.L., M.L., Ö.A., H.Y., T.Z., X.K., X.L., and A.M. tested the system and provided feedback for alignment. M.U. co-conceived the project and supervised the study. C.Z. co-conceived the project, guided research direction, supervised the study, and revised the manuscript.

Competing interests

The authors declare no competing interests.

References

1. M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Å. Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. Al-Khalili Szigyarto, J. Odeberg, D. Djureinovic, J. Ottosson Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, F. Pontén, Proteomics. Tissue-based map of the human proteome. *Science* 347, 1260419 (2015).
2. M. Uhlén, C. Zhang, S. Lee, E. Sjöstedt, L. Fagerberg, G. Bidkhori, R. Benfeitas, M. Arif, Z. Liu, F. Edfors, K. Sanli, K. von Feilitzen, P. Oksvold, E. Lundberg, S. Hober, P. Nilsson, J. Mattsson, J. M. Schwenk, H. Brunnström, B. Glimelius, T. Sjöblom, P.-H. Edqvist, D. Djureinovic, P. Micke, C. Lindskog,

A. Mardinoglu, F. Pontén, A pathology atlas of the human cancer transcriptome. *Science* 357, eaan2507 (2017).

3. M. Karlsson, C. Zhang, L. Méar, W. Zhong, A. Digre, B. Katona, E. Sjöstedt, L. Butler, J. Odeberg, P. Dusart, F. Edfors, P. Oksvold, K. von Feilitzen, M. Zwahlen, M. Arif, O. Altay, X. Li, M. Ozcan, A. Mardinoglu, L. Fagerberg, J. Mulder, Y. Luo, F. Pontén, M. Uhlén, C. Lindskog, A single-cell type transcriptomics map of human tissues. *Sci. Adv.* 7, eabh2169 (2021).
4. P. J. Thul, L. Åkesson, M. Wiking, D. Mahdessian, A. Geladaki, H. Ait Blal, T. Alm, A. Asplund, L. Björk, L. M. Breckels, A. Bäckström, F. Danielsson, L. Fagerberg, J. Fall, L. Gatto, C. Gnann, S. Hober, M. Hjelmare, F. Johansson, S. Lee, C. Lindskog, J. Mulder, C. M. Mulvey, P. Nilsson, P. Oksvold, J. Rockberg, R. Schutten, J. M. Schwenk, Å. Sivertsson, E. Sjöstedt, M. Skogs, C. Stadler, D. P. Sullivan, H. Tegel, C. Winsnes, C. Zhang, M. Zwahlen, A. Mardinoglu, F. Pontén, K. von Feilitzen, K. S. Lilley, M. Uhlén, E. Lundberg, A subcellular map of the human proteome. *Science* 356, eaal3321 (2017).
5. A. Persson, S. Hober, M. Uhlén, A human protein atlas based on antibody proteomics. *Curr. Opin. Mol. Ther.* 8, 185–190 (2006).
6. M. Uhlén, M. J. Karlsson, W. Zhong, A. Tebani, C. Pou, J. Mikes, T. Lakshmikanth, B. Forsström, F. Edfors, J. Odeberg, A. Mardinoglu, C. Zhang, K. von Feilitzen, J. Mulder, E. Sjöstedt, A. Hober, P. Oksvold, M. Zwahlen, F. Pontén, C. Lindskog, Å. Sivertsson, L. Fagerberg, P. Brodin, A genome-wide transcriptomic analysis of protein-coding genes in human blood cells. *Science* 366, eaax9198 (2019).
7. M. Uhlén, M. J. Karlsson, A. Hober, A.-S. Svensson, J. Scheffel, D. Kotol, W. Zhong, A. Tebani, L. Strandberg, F. Edfors, E. Sjöstedt, J. Mulder, A. Mardinoglu, A. Berling, S. Ekblad, M. Dannemeyer, S. Kanje, J. Rockberg, M. Lundqvist, M. Malm, A.-L. Volk, P. Nilsson, A. Månberg, T. Dodig-Crnkovic, E. Pin, M. Zwahlen, P. Oksvold, K. von Feilitzen, R. S. Häussler, M.-G. Hong, C. Lindskog, F. Pontén, B. Katona, J. Vuu, E. Lindström, J. Nielsen, J. Robinson, B. Ayoglu, D. Mahdessian, D. Sullivan, P. Thul, F. Danielsson, C. Stadler, E. Lundberg, G. Bergström, A. Gummesson, B. G. Voldborg, H. Tegel, S. Hober, B. Forsström, J. M. Schwenk, L. Fagerberg, Å. Sivertsson, The human secretome. *Sci. Signal.* 12, eaaz0274 (2019).
8. L. Fagerberg, B. M. Hallström, P. Oksvold, C. Kampf, D. Djureinovic, J. Odeberg, M. Habuka, S. Tahmasebpoor, A. Danielsson, K. Edlund, A. Asplund, E. Sjöstedt, E. Lundberg, C. Al-Khalili Szigyarto, M. Skogs, J. Ottosson Takanen, H. Berling, H. Tegel, J. Mulder, P. Nilsson, J. M. Schwenk, C. Lindskog, F. Danielsson, A. Mardinoglu, Å. Sivertsson, K. von Feilitzen, M. Forsberg, M. Zwahlen, I. Olsson, S. Navani, M. Huss, J. Nielsen, F. Pontén, M. Uhlén, Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics* 13, 397–406 (2014).
9. P. J. Thul, C. Lindskog, The human protein atlas: a spatial map of the human proteome. *Protein Sci.* 27, 233–244 (2018).
10. M. Uhlén, E. Björling, C. Agaton, C. Al-Khalili Szigyarto, B. Amini, E. Andersen, A.-C. Andersson, P. Angelidou, A. Asplund, C. Asplund, L. Berglund, K. Bergström, H. Brumer, D. Cerjan, M. Ekström, A. Elobeid, C. Eriksson, L. Fagerberg, R. Falk, J. Fall, M. Forsberg, M. G. Björklund, K. Gumbel, A. Halimi,

I. Hallin, C. Hamsten, M. Hansson, M. Hedhammar, G. Hercules, C. Kampf, K. Larsson, M. Lindskog, W. Lodewyckx, J. Lund, J. Lundeberg, K. Magnusson, E. Malm, P. Nilsson, J. Odling, P. Oksvold, I. Olsson, E. Oster, J. Ottosson, L. Paavilainen, A. Persson, R. Rimini, J. Rockberg, M. Runeson, Å. Sivertsson, A. Sköllermo, J. Steen, M. Stenvall, F. Sterky, S. Strömberg, M. Sundberg, H. Tegel, S. Tourle, E. Wahlund, A. Waldén, J. Wan, H. Wernérus, J. Westberg, K. Wester, U. Wrethagen, L. L. Xu, S. Hober, F. Pontén, A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* 4, 1920–1932 (2005).

11. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* 33, 1877–1901 (2020).
12. OpenAI, GPT-4 technical report. arXiv:2303.08774 [cs.CL] (2023).
13. Gemini Team, Gemini: a family of highly capable multimodal models. arXiv:2312.11805 [cs.CL] (2023).
14. Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, H. S. Chan, A. Madotto, P. Fung, Survey of hallucination in natural language generation. *ACM Comput. Surv.* 55, 1–38 (2023).
15. L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, T. Liu, A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. arXiv:2311.05232 [cs.CL] (2023).
16. R. Bommasani, P. Liang, T. Lee, Holistic evaluation of language models. *Ann. N. Y. Acad. Sci.* 1525, 140–146 (2023).
17. M. Baker, 1,500 scientists lift the lid on reproducibility. *Nature* 533, 452–454 (2016).
18. C. G. Begley, J. P. A. Ioannidis, Reproducibility in science: improving the standard for basic and preclinical research. *Circ. Res.* 116, 116–126 (2015).
19. L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, J.-R. Wen, A survey on large language model based autonomous agents. *Front. Comput. Sci.* 18, 186345 (2024).
20. J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, M. S. Bernstein, Generative agents: interactive simulacra of human behavior. *Proc. 36th ACM UIST*, 1–22 (2023).
21. Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, C. Wang, AutoGen: enabling next-gen LLM applications via multi-agent conversation. arXiv:2308.08155 [cs.AI] (2023).
22. J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* 35, 24824–24837 (2022).
23. T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: language models can teach themselves to use tools. *Adv. Neural Inf. Process.*

Syst. 36, 68539–68551 (2023).

24. X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models. arXiv:2203.11171 [cs.CL] (2022).
25. C. Qian, W. Liu, H. Liu, N. Chen, Y. Dang, J. Li, C. Yang, W. Chen, Y. Su, X. Cong, J. Xu, D. Li, Z. Liu, M. Sun, ChatDev: communicative agents for software development. *Proc. 62nd Annu. Meet. Assoc. Comput. Linguist.*, 15174–15186 (2024).
26. X. Deng, Y. Gu, B. Zheng, S. Chen, S. Stevens, B. Wang, H. Sun, Y. Su, Mind2Web: towards a generalist agent for the web. *Adv. Neural Inf. Process. Syst.* 36, 28091–28114 (2023).
27. S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, T. Ou, Y. Bisk, D. Fried, U. Alon, G. Neubig, WebArena: a realistic web environment for building autonomous agents. arXiv:2307.13854 [cs.AI] (2023).
28. D. A. Boiko, R. MacKnight, B. Kline, G. Gomes, Autonomous chemical research with large language models. *Nature* 624, 570–578 (2023).
29. A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, P. Schwaller, Augmenting large language models with chemistry tools. *Nat. Mach. Intell.* 6, 525–535 (2024).
30. W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, J.-R. Wen, A survey of large language models. arXiv:2303.18223 [cs.CL] (2023).
31. G. Marcus, E. Davis, Rebooting AI: Building Artificial Intelligence We Can Trust (Pantheon, 2019).
32. M. Mitchell, D. C. Krakauer, The debate over understanding in AI's large language models. *Proc. Natl. Acad. Sci. U.S.A.* 120, e2215907120 (2023).
33. D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, J. Dean, Carbon emissions and large neural network training. arXiv:2104.10350 [cs.LG] (2021).
34. J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling laws for neural language models. arXiv:2001.08361 [cs.LG] (2020).
35. B. Shneiderman, Human-Centered AI (Oxford Univ. Press, 2022).
36. G. Mialon, R. Dessì, M. Lomeli, C. Nalmpantis, R. Pasunuru, R. Raileanu, B. Rozière, T. Schick, J. Dwivedi-Yu, A. Celikyilmaz, E. Grave, Y. LeCun, T. Scialom, Augmented language models: a survey. *Trans. Mach. Learn. Res.* (2023).
37. A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, P. Clark, Self-refine: iterative refinement with self-feedback. *Adv. Neural Inf. Process. Syst.* 36, 46534–46594 (2023).
38. E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: can language models be too big? *Proc. FAccT*, 610–623 (2021).
39. R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K.

Ethayarajh, F.-F. Li, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, P. Liang, On the opportunities and risks of foundation models. arXiv:2108.07258 [cs.LG] (2021).

40. H. Nori, Y. T. Lee, S. Zhang, D. Carignan, R. Edgar, N. Fusi, N. King, J. Larson, Y. Li, W. Liu, R. Luo, S. M. McKinney, R. O. Ness, H. Poon, T. Qin, N. Usuyama, C. White, E. Horvitz, Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv:2311.16452 [cs.CL] (2023).
41. L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, I. Gabriel, Ethical and social risks of harm from language models. arXiv:2112.04359 [cs.CL] (2021).
42. N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, S. Yao, Reflexion: language agents with verbal reinforcement learning. *Adv. Neural Inf. Process. Syst.* 36, 8634–8652 (2023).
43. S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, ReAct: synergizing reasoning and acting in language models. arXiv:2210.03629 [cs.AI] (2022).
44. Significant-Gravitas, AutoGPT: build, deploy, and run AI agents (GitHub repository, accessed 29 March 2026).
45. The UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* 51, D523–D531 (2023).
46. F. Cunningham, J. E. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amode, I. M. Armean, O. Austine-Orimoloye, A. G. Azov, I. Barnes, R. Bennett, A. Berry, J. Bhai, A. Bignell, K. Billis, S. Boddu, L. Brooks, M. Charkhchi, C. Cummins, L. Da Rin Fioretto, C. Davidson, K. Dodiya, S. Donaldson, B. El Houdaigui, T. El Naboulsi, R. Fatima, C. Garcia Giron, T. Genez, J. Gonzalez Martinez, C. Guijarro-Clarke, A. Gymer, M. Hardy, Z. Hollis, T. Hourlier, T. Hunt, T. Juettemann, V. Kaikala, M. Kay, I. Lavidas, T. Le, D. Lemos, J. C. Marugán, S. Mohanan, A. Mushtaq, M. Naven, D. N. Ogeh, A. Parker, A. Parton, M. Perry, I. Piližota, I. Prosovetskaia, M. P. Sakthivel, A. I. A. Salam, B. M. Schmitt, H. Schuilenburg, D. Sheppard, J. G. Pérez-Silva, W. Stark, E. Steed, K. Sutinen, R. Sukumaran, D. Sumathipala, M.-M. Suner, M. Szpak, A. Thormann, F. F. Tricomi, D. Urbina-Gómez, A. Veidenberg, T. A. Walsh, B. Walts, N. Willhoft, A. Winterbottom, E. Wass, M. Chakiachvili, B. Flint, A. Frankish, S. Giorgetti, L. Haggerty, S. E. Hunt, G. R. IIsley, J. E. Loveland, F. J. Martin, B. Moore, J. M. Mudge, M. Muffato, E. Perry, M.

Ruffier, J. Tate, D. Thybert, S. J. Trevanion, S. Dyer, P. W. Harrison, K. L. Howe, A. D. Yates, D. R. Zerbino, P. Flicek, Ensembl 2022. *Nucleic Acids Res.* 50, D988–D995 (2022).

47. Gene Ontology Consortium, S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, P. Gaudet, N. L. Harris, D. P. Hill, R. Lee, H. Mi, S. Moxon, C. J. Mungall, A. Muruganugan, T. Mushayahama, P. W. Sternberg, P. D. Thomas, K. Van Auken, J. Ramsey, D. A. Siegele, R. L. Chisholm, P. Fey, M. C. Aspromonte, M. V. Nugnes, F. Quaglia, S. Tosatto, M. Giglio, S. Nadendla, G. Antonazzo, H. Attrill, G. dos Santos, S. Marygold, V. Strelets, C. J. Tabone, J. Thurmond, P. Zhou, S. H. Ahmed, P. Asanitthong, D. Luna Buitrago, M. N. Erdol, M. C. Gage, M. Ali Kadhum, K. Y. C. Li, M. Long, A. Michalak, A. Pesala, A. Pritazahra, S. C. C. Saverimuttu, R. Su, K. E. Thurlow, R. C. Lovering, C. Logie, S. Oliferenko, J. Blake, K. Christie, L. Corbani, M. E. Dolan, H. J. Drabkin, D. P. Hill, L. Ni, D. M. Sitnikov, M. A. Harris, S. G. Oliver, K. Rutherford, V. Wood, J. Hayles, J. Bähler, A. Lock, E. R. Bolton, J. De Pons, M. L. Dwinell, G. T. Hayman, D. E. Laulederkind, S. J. F. Laulederkind, M. R. Nigam, S. J. Wang, M. Westerfield, The Gene Ontology knowledgebase in 2023. *Genetics* 224, iyad031 (2023).
48. E. J. Topol, High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56 (2019).
49. P. Rajpurkar, E. Chen, O. Banerjee, E. J. Topol, AI in health and medicine. *Nat. Med.* 28, 31–38 (2022).
50. Y. Gil, C. H. David, I. Demir, B. T. Essawy, R. W. Fulweiler, J. L. Goodall, L. Karlstrom, H. Lee, H. J. Mills, J.-H. Oh, S. A. Pierce, A. Pope, M. W. Tzeng, S. R. Villamizar, X. Yu, Toward the geoscience paper of the future: best practices for documenting and sharing research from data to software to provenance. *Earth Space Sci.* 3, 388–415 (2016).
51. National Academies of *Science*s, Engineering, and Medicine, Reproducibility and Replicability in Science (National Academies Press, 2019).

Table

Table 1. ASO benchmark evaluation across six research queries of increasing complexity. All runs executed autonomously with no human intervention.

Difficulty	Query	Measurements	Charts	Status
Easy	Top 25 genes enriched in heart muscle in a lollipop chart	15	1	Pass
Easy	Show me ALB/HP/CYP3A4 liver vs kidney in a grouped bar chart	35	1	Pass
Medium	Show me metabolic enzymes expression difference liver vs pancreas as scatter plot	534	1	Pass
Medium	Find receptors enriched in cerebral cortex. Lollipop bottom 20, grouped bar cortex vs liver top 25	106	2	Pass
Medium	Find TFs enriched in heart vs brain. Lollipop each, scatter head-to-head	88	3	Pass
Hard	Find transcription factors enriched in liver. Measure in liver and kidney. Lollipop top 20, grouped bar top 15, scatter, heatmap top 10	228	4	Pass*
Total		1,006	12	100%

Pass* = guardrail-mediated self-correction. Code guardrail rejected ranked data passed to scatter tool; orchestrator autonomously re-routed through analyze_scatter with raw measurement data.

Figures

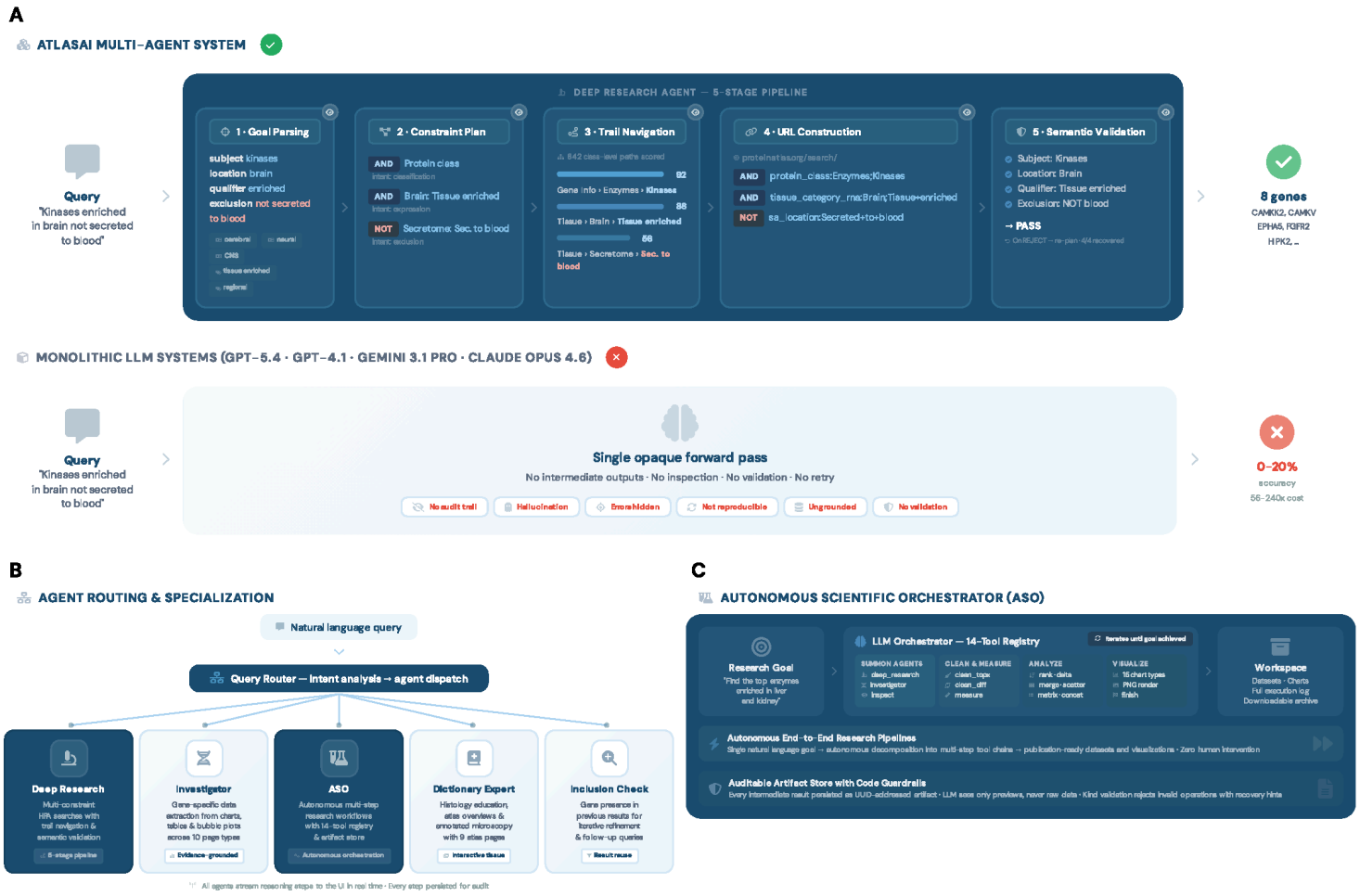


Figure 1

Multi-agent architecture for Human Protein Atlas navigation. (A) Comparison of AtlasAI multi-agent pipeline (top) showing five explicit processing stages with auditable intermediate outputs versus monolithic LLM approach (bottom) where reasoning is opaque and produces 0 to 20% accuracy at 56 to 240 times higher cost. (B) Agent routing hub: natural language query is analyzed and dispatched to one of five specialized agents (Deep Research, Investigator, ASO, Dictionary Expert, Inclusion Check), with routing decisions logged. (C) ASO overview: natural language goal enters LLM orchestrator with 14-tool registry, iterates through autonomous execution, persists all artifacts in auditable workspace.

A

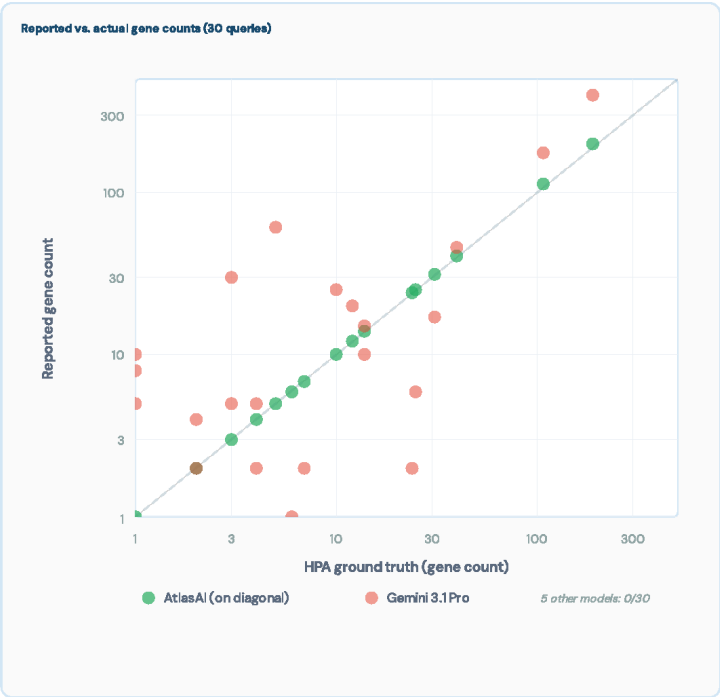
Benchmark comparison: AtlasAI vs. frontier language models on 13 representative HPA queries. Correctness: ±10%. All models given web browsing. Full 30-query results in Table S3.

QUERY	TRUTH	ATLASAI GPT-4.1 mini	CLAUDE 4.6 Extended	GEMINI 3.1 PRO Thinking	GPT-5.4 High	GPT-5.4 Medium	GPT-5.4 None	GPT-4.1 None
EASY (1–2 CONSTRAINTS)								
Proteins specific to placenta, secreted to blood	10	10 ✓	Refused	~25 ✗	Refused	Refused	Refused	Refused
Metabolic proteins enriched only in liver	107	107 ✓	Refused	~175 ✗	Refused	Refused	Refused	Refused
MEDIUM (3 CONSTRAINTS)								
Nuclear proteins enriched in testis, not in liver	188	188 ✓	Refused	~400 ✗	Refused	Refused	Refused	Refused
Cardiomyocyte enzymes, not secreted to blood	3	3 ✓	Refused	~30 ✗	Refused	Refused	Refused	Refused
Proteases in pancreas, secreted to digestive system	14	14 ✓	Refused	~15 ✓	Refused	Refused	Refused	Refused
Enzymes in hypothalamus, not secreted to blood	1	1 ✓	Refused	≤5 ✗	Refused	Refused	Refused	Refused
HARD (4 CONSTRAINTS)								
CD markers in T-cells, not B-cells or NK-cells	25	25 ✓	Refused	6 ✗	Refused	Refused	Refused	Refused
Cancer-related metabolic proteins in liver, not kidney	7	7 ✓	Refused	~2 ✗	Refused	Refused	Refused	Refused
Metabolic enzymes in cerebellum, not in cortex	6	6 ✓	Refused	0 ✗	Refused	Refused	Refused	Refused
FDA membrane proteins in brain, not secreted	40	40 ✓	Refused	~40–50 ✓	Refused	Refused	Refused	Refused
VERY HARD (5–6 CONSTRAINTS)								
TFs in testis, not brain/liver/ovary/prostate	24	24 ✓	Refused	1–3 ✗	Refused	Refused	Refused	Refused
Kinases in brain, not liver/heart/kidney, not secreted	4	4 ✓	Refused	~5 ✗	Refused	Refused	Refused	Refused
FDA targets in brain, not liver/kidney/heart, not secreted	31	31 ✓	Refused	~17 ✗	Refused	Refused	Refused	Refused
ACCURACY		100%	0%	20%	0%	0%	0%	0%
ATLAS COST INDEX		1.0x	240x	86x	124x	78x	56x	60x

Correct (±10%) Wrong Refused

ACI = API cost relative to AtlasAI (1.0x). All models given web browsing access.

B



C

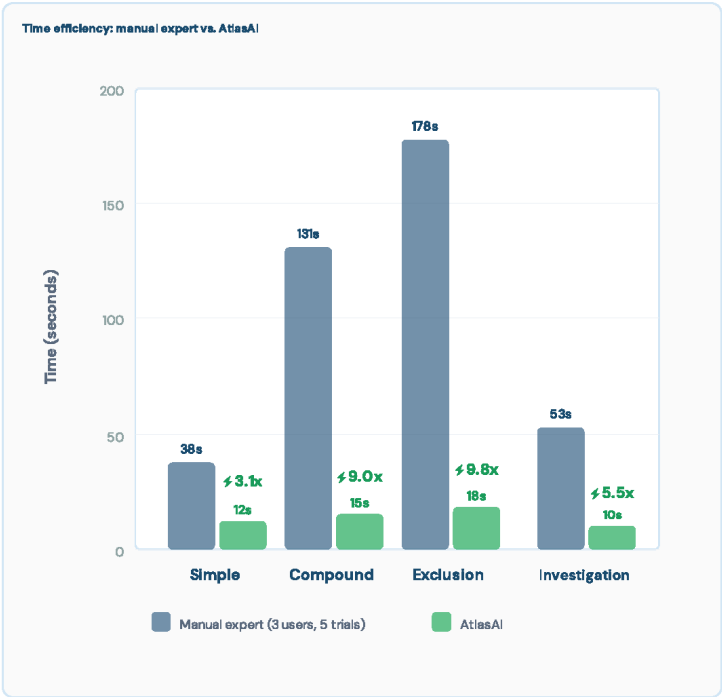


Figure 2

Benchmark performance against frontier language models. (A) Per-query comparison table: AtlasAI achieves complete task resolution across 13 representative queries (full 30-query results in Table S3) while Claude Opus 4.6, GPT-5.4 (three reasoning levels), GPT-4.1, and Gemini 3.1 Pro achieve 0 to 20%. Atlas Cost Index shows 56 to 240 times higher cost for frontier models. (B) Reported versus actual gene counts for all 30 queries on log scale. AtlasAI values fall on the identity diagonal; Gemini 3.1 Pro values are scattered across the plot; five other models produced zero data points. (C) Time efficiency: manual

expert navigation versus AtlasAI across four query categories, showing 3.1 to 9.8 fold speedup that increases with query complexity.

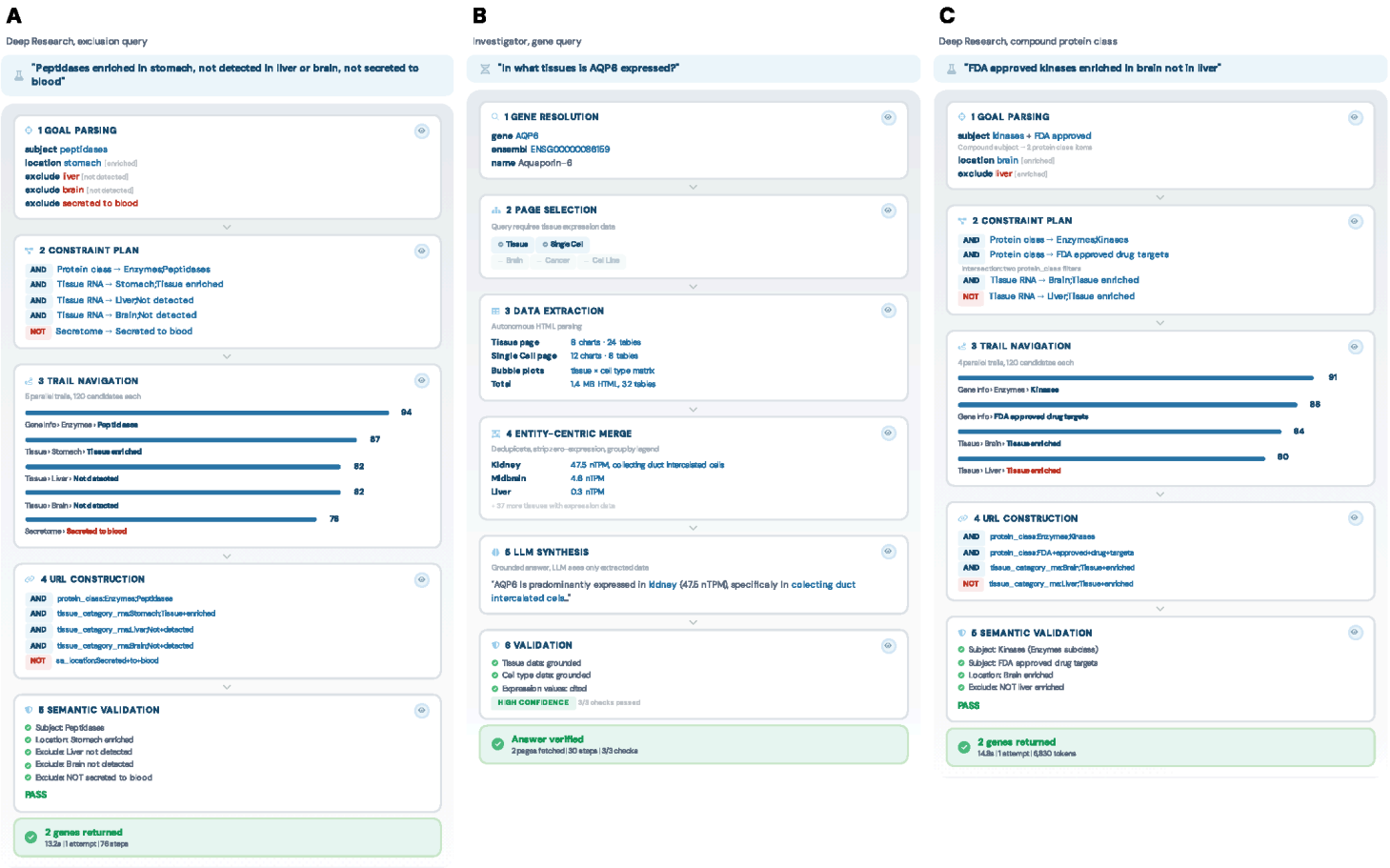


Figure 3

Transparent reasoning traces for three query types. (A) Deep Research exclusion query: "Peptidases enriched in stomach, not detected in liver or brain, not secreted to blood." Five-stage pipeline showing goal parsing, constraint planning with AND/NOT operators, trail navigation with hybrid scoring, URL construction, and semantic validation. 2 genes returned, 13.2 seconds. (B) Investigator gene query: "In what tissues is AQP6 expressed?" Six-stage trace showing gene resolution, page selection, autonomous data extraction (8 charts, 24 tables), entity-centric merge, LLM synthesis from grounded evidence, and validation. (C) Deep Research compound protein class: "FDA approved kinases enriched in brain not in liver." Shows compound subject decomposition into two protein class filters, parallel trail navigation, intersection URL construction, and validation.

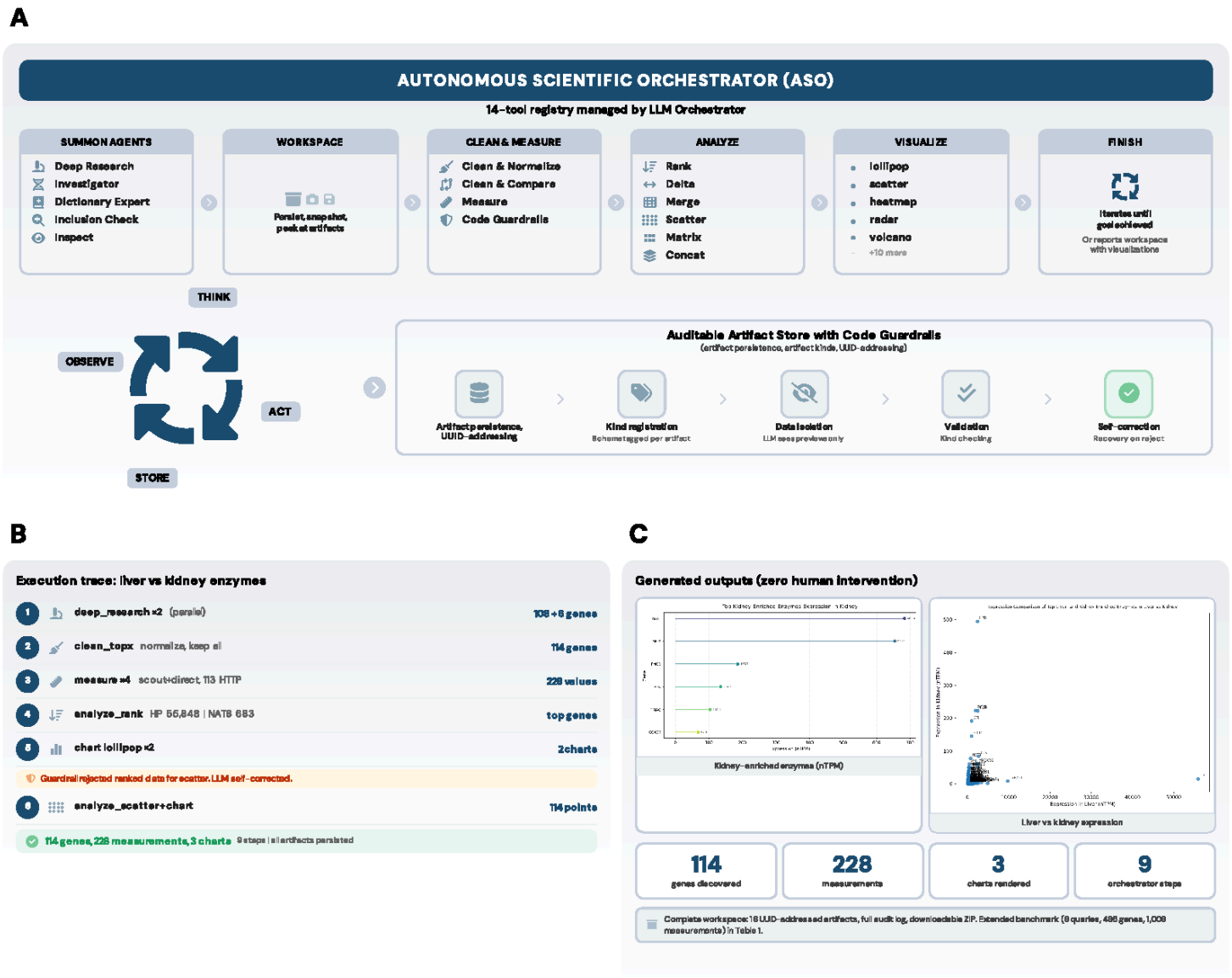


Figure 4

Autonomous Scientific Orchestrator (ASO) architecture and execution. (A) System overview: 14-tool registry organized into five categories (Summon Agents, Workspace, Clean and Measure, Analyze, Visualize, Finish), orchestrated through a think-act-store-observe loop. The auditable artifact store provides UUID-addressed persistence, kind registration, data isolation (LLM sees previews only), validation, and self-correction on reject. (B) Execution trace for the liver versus kidney enzyme comparison query, showing 9 orchestrator steps from parallel deep research through scout-and-direct measurement (1 LLM call, 113 HTTP fetches), ranking (HP 55,848 nTPM, NAT8 683 nTPM), and chart generation, including guardrail-mediated self-correction for the scatter plot. (C) Generated outputs from the same run: kidney-enriched enzyme lollipop chart and liver versus kidney expression scatter plot. 114 genes, 228 measurements, 3 charts, 9 steps, all artifacts persisted in downloadable workspace.

Supplementary Files

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

- supplementarymaterial.pdf
- AtlasAI91.mp4