**Knowledge Graph and Chain-of-Thought Enhanced Data Mining for Multi-modality Neuronal Neuroscience**

**Abstract**

Neuroscience still lacks a unified multimodal resource that systematically integrates neuron morphology, projection, and transcriptomic data. To fill this gap, we developed a large-scale database combining two state-of-the-art brain atlases, 294 brain regions and 923 subregions, 182,483 neurons with reconstructed 3-D arbors, and 1,122 genes profiled across 5.24 million cells. On this foundation, we built **NeuroXiv Knowledge Graph (NeuroXiv-KG)**, a knowledge graph that encodes 34,771 nodes that correspond to brain regions, subregions, neurons and transcriptomic cell types, and 252.6 million cross-modality relationships, capturing the complex associations among molecular, morphological, and neuronal connection domains. Further, we introduced AI Powered Open Mining with Chain of Thought (**AIPOM-CoT)**, a schema-adaptive chain of thought agent that converts natural language prompts into multi-step analytical workflows involving graph retrieval, statistics, and provenance tracking. AIPOM-CoT can interpret a biologist’s question, execute multi-stage reasoning, and return automated and reproducible results. We demonstrate its performance through two applications: (1) an analysis of Car3-positive neurons that identifies their subclasses, anatomical localization, projection networks, and molecular fingerprints; and (2) a whole-brain tri-modal fingerprint map that links molecular, morphological, and projection profiles and systematically ranks pairwise agreement and mismatch—revealing where molecular patterns does or does not predict morphology or connectivity. Together, NeuroXiv-KG, and AIPOM-CoT provide a scalable AI platform for cross-modality reasoning, accelerating new discovery of neuroscience.

**Introduction**

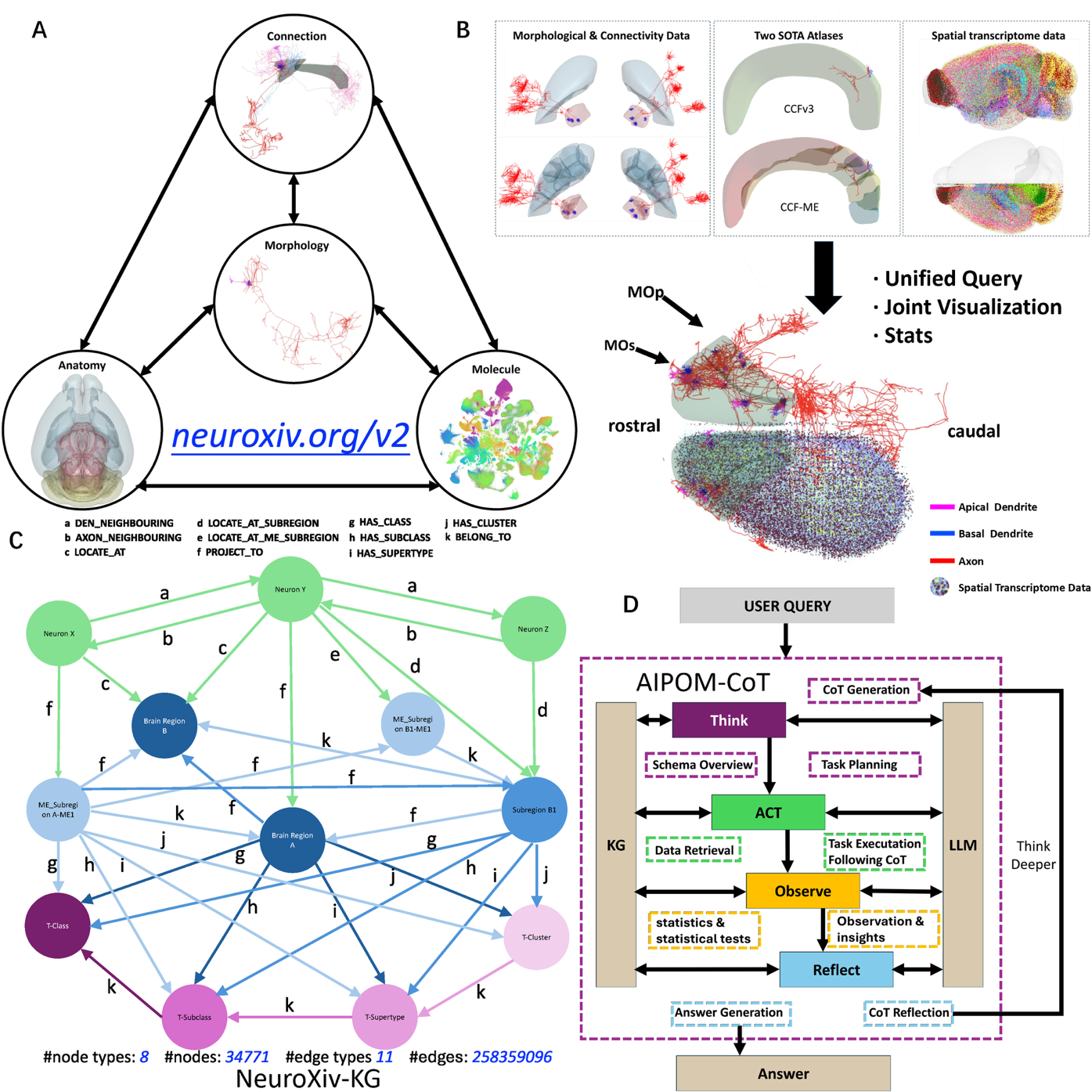
Over the past decade, major mouse-brain datasets have expanded across modalities. Common coordinate frameworks such as the Allen Mouse Brain CCFv3 provide a population-average anatomical reference spanning hundreds of regions [1]; mesoscale projection resources and serial two-photon tomography measure long-range wiring under standardized protocols [2,3]; large collections of single-neuron reconstructions now cover broad brain territories and reveal both stereotypy and diversity across putative transcriptomic classes [6]; and whole-brain spatial transcriptomics places thousands of transcriptomic types into anatomical space [4,5]. Our earlier NeuroXiv 1.0 aggregated parts of these streams to facilitate exploratory analyses at scale. Yet, despite this progress, these resources largely live in separate silos—with heterogeneous schemas, coordinate systems, and access patterns—so that unified cross-modal integration at the region level, and especially at single-neuron granularity, remains difficult.

Spatial granularity further complicates synthesis. The Mouse Brain Atlas of Dendritic Microenvironments (CCF-ME) subdivides parcels using local dendritic context (>100k neurons), improving anatomical discrimination and correlating with projection specificity while remaining compatible with CCF space [8]. Together, CCFv3 (population reference) and CCF-ME (morphology-informed fine parcellation) provide complementary coordinates on which morphology, projections, and spatial transcriptomics could, in principle, be analyzed jointly [1,8]. What has been missing is a large, unified multimodal database that systematically brings these modalities into one substrate and exposes explicit cross-links (membership, projection, composition) so that region-level questions can be asked and answered directly.

Within this landscape, existing ecosystems each address important pieces. The Allen Brain Knowledge Portal / Cell Type Knowledge Explorer curate high-quality cell-type resources with excellent browsing [22]. Blue Brain Nexus offers a powerful, generic KG/data-management backbone (RDF/ontologies) focused on FAIR modeling and versioning [23]. The BrainGlobe suite streamlines 3D imaging workflows for detection, registration, atlas mapping, and visualization [9,10]. Beyond mouse neuroanatomy, human meta-analytic text-to-map tools (NeuroSynth, NeuroQuery) automate language-to-map associations [17,18], and generic LLM-agent frameworks (ReAct, Toolformer, AutoGPT, LangChain/LangGraph) demonstrate planning and tool use [15,16,19–21]. Taken together, however, none of these provide a single, region-level system that unifies morphology, projections, and spatial transcriptomics in one graph and turns natural-language questions into audited, reproducible analyses.

To fill this gap, we assembled a large-scale database anchored to two state-of-the-art atlases (CCFv3 and CCF-ME), comprising 294 brain regions and 923 subregions, 182,483 neurons with reconstructed 3-D arbors, and 1,122 genes profiled across 5.24 million cells. On this foundation we built the NeuroXiv Knowledge Graph (NeuroXiv-KG), which encodes 34,771 nodes (anatomical parcels and ME subparcels, reconstructed neurons, and transcriptomic tiers—Class/Subclass/Supertype/Cluster) and 252.6 million cross-modality relationships capturing explicit membership, projection, and composition links [1,4,8]. We then introduce AI-Powered Open Mining with Chain of Thought (AIPOM-CoT), a schema-adaptive agent that compiles a biologist’s prompt into multi-step analytical workflows involving graph traversal and statistical estimation (effect sizes, confidence intervals, permutation-based p with FDR correction), with full provenance capture (snapshot/seed, tool list, parameters, inputs/outputs). We demonstrate two representative applications: a live Car3-positive analysis that identifies subclasses, localizes anatomical pockets, maps projection networks, and profiles molecular fingerprints of repeatedly hit targets; and a global fingerprint survey that assigns tri-modal (molecular, morphology, projection) fingerprints to every region to derive similarity structure and divergence (mismatch) pairs. Together, NeuroXiv-KG and AIPOM-CoT provide the unified multimodal resource and analysis capability needed to move from disparate datasets to transparent, reproducible, and testable cross-modal insight.

**Results**

**Integrating multi-modal neuroscience data into a unified knowledge graph with AI-powered automated mining**

**Figure 1 |** *Integrating multi-modal neuroscience data into a unified knowledge graph with AI-powered automated mining.* (A) Conceptual framework for unifying four complementary data modalities—anatomical structure, cellular morphology, axonal connectivity, and molecular composition—into a single queryable knowledge graph. These modalities provide different perspectives on brain organization: anatomy defines spatial boundaries and hierarchical relationships; morphology captures dendritic and axonal arbor geometry; connectivity maps circuit wiring through axonal projections; and molecular data reveals cellular identity through transcriptomic profiles. Traditional neuroscience workflows require navigating separate databases and manually integrating across modalities. Our system unifies these data types within a structured knowledge representation, enabling automated cross-modal queries and systematic comparative analyses that were previously intractable. (B) Comprehensive data integration pipeline. Top panels show representative examples of single-neuron morphological reconstructions paired with connectivity matrices, curated from multiple sources including the Allen Brain Atlas, MouseLight, and other community repositories. Middle panels display two state-of-the-art 3D anatomical reference atlases—CCFv3 (Allen Common Coordinate Framework version 3) and CCF-ME (merged embryonic extension)—providing spatial registration frameworks across developmental stages and ensuring anatomical consistency. Bottom panels illustrate high-resolution spatial transcriptomic data from MERFISH (multiplexed error-robust fluorescence in situ hybridization), capturing single-cell molecular profiles with spatial coordinates. The integrated platform harmonizes morphological reconstructions (~34,000 neurons), connectivity matrices, anatomical parcellations (~300 brain regions), and spatial transcriptomic datasets (>4.5 million cells) within a unified coordinate system, creating a multi-scale representation spanning molecular, cellular, and systems levels. (C) NeuroXiv-KG schema and scale. The knowledge graph comprises 8 node types representing biological entities (neurons, brain regions, cell types, genes, etc.) and 11 edge types encoding relationships (neuronal projections, regional containment, gene expression, morphological features, etc.). The current instantiation contains 34,771 nodes interconnected by over 258 million edges, forming a richly connected semantic network. Colored nodes illustrate different entity classes: anatomical regions (blue), neurons (green), molecular markers (purple), and cell-type clusters (pink). Edge types include both explicitly asserted relationships (e.g., "LOCATE\_AT" connecting neurons to brain regions, "PROJECT\_TO" encoding axonal projections) and computationally derived links (e.g., "HAS\_CLASS" from cell-type clustering, "BELONG\_TO" for hierarchical containment). This structured representation transforms heterogeneous neuroscience data into a machine-readable format amenable to automated reasoning and systematic discovery (D) AIPOM-CoT agent architecture for schema-adaptive automated analysis. The agent employs a cognitive loop integrating large language models (LLMs) with knowledge graph operations. Upon receiving a user query in natural language, the \*\*Think\*\* module generates a chain-of-thought reasoning plan, decomposing complex questions into executable sub-tasks while dynamically selecting appropriate knowledge graph schemas and query patterns. The \*\*ACT\*\* module translates planned steps into concrete knowledge graph retrievals and computational operations, accessing the unified data repository. The \*\*Observe\*\* module evaluates retrieved results, performing statistical analyses and extracting insights from multi-modal data. The \*\*Reflect\*\* module assesses progress toward answering the original query, identifies gaps or inconsistencies, and generates follow-up reasoning steps, creating an iterative refinement loop. This architecture enables the system to handle open-ended biological questions without pre-programmed workflows, automatically determining which data modalities to query, how to integrate them, and when the assembled evidence sufficiently addresses the question. Critically, all reasoning steps, data retrievals, and computational operations are logged with full provenance, ensuring reproducibility and enabling users to validate or refine the automated analysis

Modern neuroscience generates data across multiple complementary scales and modalities: spatial transcriptomics reveals molecular cell-type identity at single-cell resolution, morphological reconstructions capture dendritic and axonal arbor geometry, connectivity atlases map circuit wiring through projection patterns, and anatomical reference frameworks provide hierarchical spatial organization. Each modality offers a distinct lens on brain organization, yet they reside in separate repositories with heterogeneous formats, annotations, and coordinate systems. Integrating across these modalities remains a manual, labor-intensive process requiring expert knowledge of multiple databases, custom data parsing scripts, and ad hoc procedures for cross-referencing entities—a workflow that scales poorly as datasets grow and becomes a bottleneck for hypothesis generation and systematic comparative analyses. Despite the availability of rich data resources, the lack of unified, machine-readable integration has confined most neuroscience analyses to single-modality studies or small-scale manual integration efforts, leaving cross-modal patterns largely unexplored.

To address this integration challenge, we developed NeuroXiv-KG, a comprehensive knowledge graph unifying four fundamental data modalities—molecular composition, cellular morphology, circuit connectivity, and anatomical organization—within a single structured semantic network (Figure 1A-C). The knowledge graph integrates single-neuron morphological reconstructions from multiple sources including MouseLight and the Allen Brain Atlas (~34,000 neurons with detailed dendritic and axonal arbors), spatial transcriptomic datasets from MERFISH capturing molecular profiles of >4.5 million cells across the mouse brain, axonal projection connectivity matrices spanning hundreds of brain regions, and 3D anatomical reference atlases (CCFv3 and CCF-ME) providing spatial registration across developmental stages and ensuring anatomical consistency (Figure 1B).

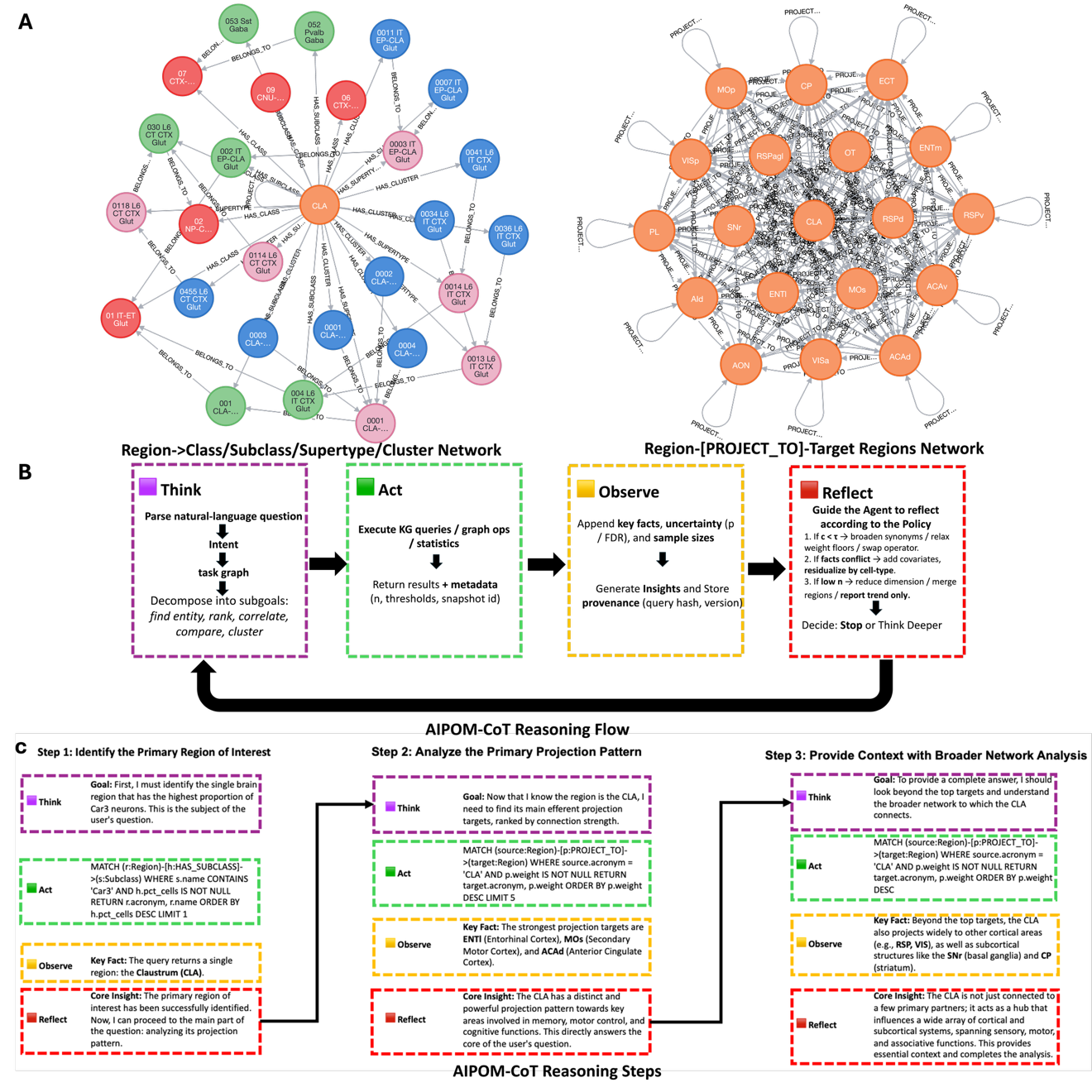
We harmonized these heterogeneous data types within a unified coordinate framework and structured them using a formal schema comprising 8 node types (representing biological entities such as neurons, brain regions, cell types, genes, and morphological features) and 11 edge types (encoding relationships including neuronal projections, regional containment, gene expression, and morphological properties) (Figure 1C). The resulting knowledge graph contains 34,771 nodes interconnected by over 258 million edges, forming a richly connected semantic network where queries can traverse across modalities—for example, starting from a molecular marker, identifying enriched cell types, locating brain regions, retrieving morphological reconstructions, and mapping their projection targets. This structured representation transforms fragmented neuroscience datasets into a unified, machine-readable resource amenable to systematic exploration and automated reasoning.

While knowledge graphs provide unified data representation, extracting meaningful insights still requires expertise in graph query languages and domain knowledge to formulate appropriate questions. To enable automated, natural-language-driven analysis, we developed AIPOM-CoT (AI-Powered Ontology Mapping with Chain-of-Thought), a reasoning agent that translates open-ended biological questions into multi-step analysis workflows (Figure 1D). The agent employs a cognitive architecture integrating large language models with knowledge graph operations through four interconnected modules: (1) Think — generates chain-of-thought reasoning plans that decompose complex queries into executable sub-tasks while dynamically selecting appropriate knowledge graph schemas and query patterns; (2) ACT — translates planned steps into concrete knowledge graph retrievals, statistical computations, and multi-modal data integration operations; (3) Observe — evaluates retrieved results, extracts quantitative patterns, and assesses the biological significance of findings; (4) Reflect — determines whether accumulated evidence sufficiently addresses the original question, identifies gaps or inconsistencies, and generates follow-up reasoning steps, creating an iterative refinement loop.

This schema-adaptive architecture enables AIPOM-CoT to handle diverse question types without pre-programmed workflows. Given a query such as "What can you tell me about Car3+ neurons?", the agent autonomously determines the relevant data modalities (transcriptomic profiles to identify Car3-expressing cell types, regional distributions to locate enrichment, morphological data to characterize structure, projection data to map connectivity), formulates appropriate knowledge graph queries for each modality, integrates the retrieved information across scales, and generates interpretable summaries with full provenance. Critically, all reasoning steps, data retrievals, and computational operations are logged, enabling users to validate automated analyses, adjust parameters, or extend workflows—transforming atlas interaction from manual browsing of pre-computed views to planned, reproducible, multi-modal investigations.

Together, NeuroXiv-KG and AIPOM-CoT constitute a unified platform for automated cross-modal neuroscience data mining at scale. The knowledge graph provides comprehensive multi-modal data integration, while the reasoning agent enables natural-language-driven, automated analysis workflows that adapt to the structure and semantics of the underlying data. To demonstrate the system's capabilities, we present two complementary applications: First, we show how AIPOM-CoT automatically retrieves and integrates molecular, morphological, and connectivity data in response to an open-ended query about Car3+ neurons, assembling a comprehensive multi-modal neuronal profile without manual data curation or pre-specified analysis scripts (Result 3). Second, we demonstrate systematic whole-brain analysis, where the agent automatically constructs tri-modal fingerprints across all major brain regions and discovers pervasive cross-modal divergence patterns—revealing that approximately 40% of region pairs display molecular-morphological or molecular-projection mismatches, suggesting semi-independent organizational principles operating across different biological scales (Result 4).

These demonstrations illustrate how the integrated platform transforms neuroscience atlas interaction from manual, single-modality browsing to automated, cross-modal discovery. The system enables researchers to pose biological questions in natural language and receive integrated, multi-scale answers spanning molecular identity, cellular morphology, circuit connectivity, and anatomical context—analyses that would require hours to days of manual effort using conventional approaches. By providing machine-readable integration, automated reasoning capabilities, and full provenance tracking, the platform establishes a foundation for systematic, reproducible cross-modal neuroscience at scale.

**Inside AIPOM-CoT: how natural-language questions become auditable analyses**

**Figure 2 |** *AIPOM-CoT: a schema-adaptive, evidence-seeking agent built on NeuroXiv-KG.* (A) Operator-ready computation surfaces. Two live, queryable abstractions derived from the KG: a Region→Class/Subclass/Supertype/Cluster neighborhood exposing typed taxonomy edges, and a Region–[PROJECT\_TO]–Target egonet exposing directed projection weights and provenance. These views are the substrates the agent traverses and aggregates—no bespoke scripts. (B) Reasoning loop with provenance. From a natural-language prompt the agent runs a four-stage loop: Think—parse intent, build a task graph, inspect the KG schema, and choose operators (traversal, aggregation, ranking, enrichment, correlation/partial correlation, similarity, permutation tests, FDR); Act—bind operators to the schema and compile to graph queries, execute with explicit parameters and compute budgets, and return results plus metadata (n, thresholds, snapshot/seed, query hash); Observe—append effect sizes, confidence intervals, permutation-based p and FDR-adjusted q to an evidence buffer, perform coverage/stability checks, and emit intermediate insights; Reflect—apply policy rules (add covariates, sweep thresholds, switch metrics, expand context via “Think Deeper,” or stop) until halting criteria are met. (C) Step templates. Three generic recipes illustrate how plans are assembled: (1) ROI selection (rank candidate regions/taxa under coverage constraints), (2) pattern profiling (e.g., projection or morphology summaries with uncertainty calibration), and (3) context expansion & controls (neighbor comparisons, confound checks, metric stability sweeps). Each template records inputs, operator choices, parameters, outputs, and provenance, enabling replay.

Our goal is to make cross-modality reasoning executable and reproducible. AIPOM-CoT achieves this by coupling schema-aware planning with operator binding, evidence tracking, and policy-driven reflection (Fig. 2).

Schema introspection and operator binding. Given a prompt, the agent first parses intent and constructs a task graph. It inspects the live KG schema (node/edge types and attributes such as PROJECT\_TO, HAS\_CLASS/SUBCLASS/SUPERTYPE/CLUSTER, LOCATE\_AT, morphology features) and binds operators from a fixed library: graph traversals and aggregations; ranking and enrichment; correlation and partial correlation; similarity on fingerprints (molecular, morphology, projection); permutation tests with FDR control; and visualization primitives. Binding produces parameterized queries (e.g., region scopes, tier pooling, target cutoffs) compiled to graph operations.

Execution kernel and evidence buffer. During Act, the kernel runs the bound queries under explicit compute budgets, returning results with sample sizes, thresholds, and a snapshot/seed tagged by a query hash. In Observe, results are written to an evidence buffer that stores numerical estimates (effect sizes, CIs, permutation p, FDR q), coverage diagnostics, and the exact inputs/outputs for each step. This buffer powers both intermediate reasoning and later replay.

Reflection policy and halting. A policy layer analyzes the evidence buffer. If coverage is shallow, the agent pools tiers or expands the search radius; if composition confounds are detected, it adds covariates/partial correlations; if metrics are unstable, it sweeps thresholds or switches similarity metrics; if context is insufficient, it invokes a “Think Deeper” pass to broaden the plan. The loop halts when stability, coverage, and consistency criteria are satisfied or a budget limit is reached, at which point the agent emits a consolidated answer and the complete execution trace.

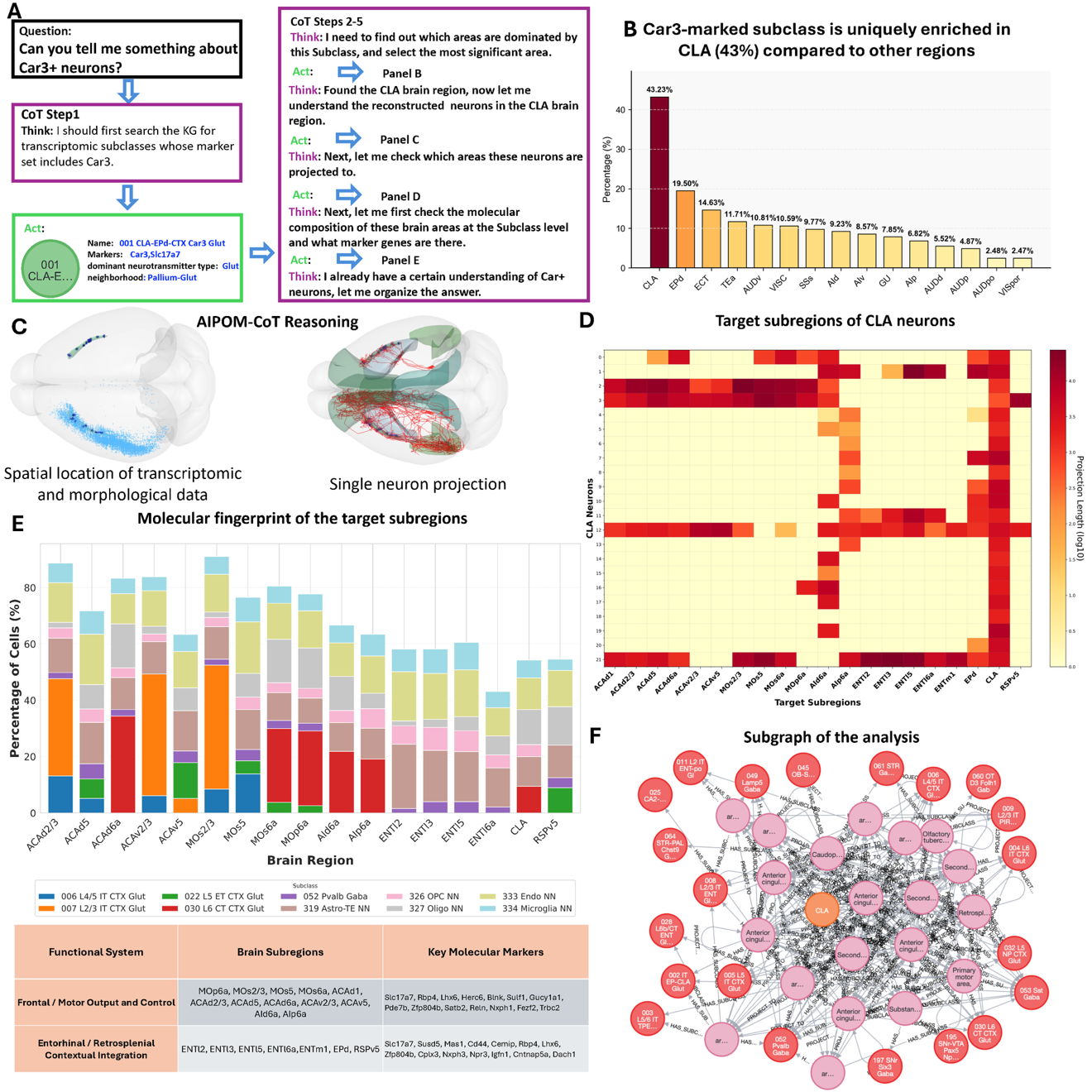
Reproducibility and auditability. Every run returns (i) a natural-language answer grounded in calibrated statistics and (ii) a machine-readable trace (operator list, parameters, inputs/outputs, snapshot/seed, query hashes). This design enables bit-for-bit replay, independent inspection, and straightforward comparison across runs or datasets.

Performance and extensibility. Typical multi-step analyses execute in a very short time end-to-end. Because operators are bound to typed relations, adding a new modality or attribute (e.g., additional morphology metrics or a new atlas split) requires only schema exposure; the same planning and reflection machinery applies without custom code.

Together, these components turn the KG into an operator-ready substrate and the natural-language prompt into an auditable workflow—a foundation we leverage in downstream results for concrete biological case studies and global, tri-modal surveys.

**Automated cross-modal data retrieval and integration for comprehensive neuronal profiling**

**Natural language queries automatically assemble multi-modality profiles of Car3+ claustrum neurons**

****

**Figure 3 |** *Natural language queries automatically assemble multi-modality profiles of Car3+ claustrum neurons.* (A) Multi-step execution plan generated by AIPOM-CoT. Given the prompt "Can you tell me something about Car3+ neurons?", the agent automatically decomposes the query into a sequence of executable steps: (i) search the knowledge graph for transcriptomic subclasses with Car3 as a marker; (ii) rank brain regions by Car3-subclass enrichment; (iii) retrieve morphological reconstructions from the top-ranked region; (iv) analyze projection targets; and (v) profile molecular composition of targets. Each step in the CoT (Chain-of-Thought) panel shows the Think → Act cycle, illustrating how natural language is translated into concrete graph operations without manual scripting. This automatic task decomposition is the foundation of the retrieval-to-integration workflow. (B) Automated regional enrichment analysis. Hypergeometric ranking across all brain regions identifies the claustrum (CLA) as uniquely enriched for Car3-marked subclasses, accounting for ~43% of occurrences—far exceeding other regions (e.g., ACAd ~19%, MOs ~14%). Importantly, CLA was not pre-selected; the system discovered this enrichment pattern automatically through statistical ranking. Bars show the percentage of Car3+ subclass cells attributed to each region. (C) Automated spatial integration of multi-modal data. Left: Spatial distribution of transcriptomic cells expressing Car3-related markers, co-registered to atlas coordinates. Right: Representative single-neuron morphological reconstruction from CLA showing soma location (blue) and long-range axonal arbor (red). The system automatically retrieves and aligns these distinct data modalities in a common coordinate frame, enabling direct comparison between molecular and structural features. This panel demonstrates automated spatial integration without manual data curation. (D) Automated aggregation of projection targets. Heat map shows a neuron-by-target matrix compiled from all available CLA reconstructions, revealing repeatedly hit downstream regions including ENTl, layer-specific cortical sites (MOs/ACAd L2/3, L5), and insular cortex (AI). Color intensity indicates relative projection strength. The system automatically aggregates axonal endpoints across neurons to identify consistent projection patterns—target discovery is data-driven rather than hypothesis-driven. (E) Automated molecular fingerprinting of target regions. Stacked bars show the transcriptomic composition (Class/Subclass/Supertype/Cluster tiers) of the repeatedly hit targets identified in (D). Cortical targets (MOs/MOp, ACAd, AI) are dominated by IT-excitatory types (L2/3, L4/5 classes, green/blue), while entorhinal/retrosplenial sites (ENTl, ENTm, RSP) show mixed IT signatures. The table groups targets into two functional systems: frontal/motor output and control, and entorhinal–retrosplenial contextual integration, with representative molecular markers listed. This automated molecular profiling reveals functional module organization as a byproduct of the technical demonstration, suggesting that CLA serves as an associative hub linking motor control with memory-related networks. (F) Provenance subgraph for replay and validation. Subgraph extracted from NeuroXiv-KG showing all nodes (regions, Car3-marked subclasses, targets) and edges (enrichment, projection, molecular composition) accessed during the workflow. The complete execution trace—including query strings, thresholds, sample sizes, snapshot/seed, and operator sequences—enables bit-for-bit replay and independent inspection. This provenance tracking ensures that the automated workflow is auditable and reproducible.

To demonstrate the system's capacity for automated cross-modal analysis, we posed an open-ended biological question to AIPOM-CoT: "Can you tell me something about Car3+ neurons?" (Figure 3A). Without manual data curation or pre-selected datasets, the agent autonomously decomposed this query into a multi-step analysis plan: (1) identify transcriptomic subclasses enriched for Car3 expression, (2) determine which brain regions show the highest enrichment for these subclasses, (3) retrieve morphological reconstructions from the enriched region, (4) map projection targets of reconstructed neurons, and (5) characterize the molecular composition of target regions. This automatic task decomposition—translating a simple natural language question into a structured, executable workflow—represents the first demonstration of on-demand cross-modal data integration in neuroscience, where the system determines what information to retrieve and how to combine it, rather than following pre-specified analysis scripts.

AIPOM-CoT first queried the knowledge graph to identify transcriptomic subclasses expressing Car3, discovering the "003 CLA-EPd-CTX Car3 Glut" subclass as the primary Car3-expressing population. The agent then automatically ranked brain regions by their enrichment for this subclass, revealing that the claustrum (CLA) contains approximately 43% of cells from this subclass—substantially higher than any other region examined (Figure 3B). Notably, the system identified this CLA enrichment through automated retrieval and quantification across the entire knowledge graph, not through hypothesis-driven selection or manual literature review. This demonstrates the system's ability to discover region-specific cellular enrichments on demand, assembling quantitative profiles from integrated molecular datasets without requiring users to know in advance where interesting patterns might emerge.

Having identified CLA as the region of maximal Car3+ subclass enrichment, AIPOM-CoT next retrieved morphological reconstruction data for CLA neurons from the knowledge graph (Figure 3C, left). The system then automatically extracted projection information from these reconstructions, identifying the target subregions innervated by CLA neurons (Figure 3C, right; Figure 3D). The projection analysis revealed a complex, heterogeneous pattern: CLA neurons target multiple cortical areas with varying strengths, including frontal, motor, and sensory regions, as well as select subcortical structures. This automated retrieval-to-integration workflow—moving seamlessly from transcriptomic identity to regional localization to morphological data to projection mapping—occurs entirely through natural language interaction, without requiring the user to manually navigate different data modalities, specify join operations, or write custom analysis code.

To complete the multi-modality profile, AIPOM-CoT automatically generated molecular fingerprints for each CLA projection target by querying cell type composition across target regions (Figure 3E). This analysis revealed the molecular diversity of the areas receiving CLA input, with different targets displaying distinct cellular compositions—some dominated by specific cortical layer markers, others by particular GABAergic subtypes. The system organized this information into an interpretable summary, including a knowledge graph visualization showing the relationships among Car3+ neurons, their projection targets, and the molecular characteristics of those targets (Figure 3F). Critically, all steps—from initial query to final integration—were executed automatically with full provenance tracking, recording the specific knowledge graph queries, data sources, and computational parameters used at each stage. This enables independent validation and iterative refinement of the analysis.

While the primary contribution of this demonstration is the automated retrieval-to-integration capability itself, the analysis as a byproduct generated an integrated multi-modal profile of Car3+ CLA neurons. The substantial CLA enrichment of Car3+ cells, combined with their widespread cortical and selective subcortical projections, is suggestive of a role in broad cortical coordination and cross-modal integration—functions previously hypothesized for the claustrum. The molecular heterogeneity of CLA target regions further suggests that Car3+ neurons may differentially modulate distinct cortical processing streams. However, testing these hypotheses would require targeted experimental validation beyond the scope of automated knowledge graph analysis.

The key advance demonstrated here is not the specific biological findings about Car3+ neurons, but rather the system's ability to assemble comprehensive multi-modal neuronal profiles on demand, in response to arbitrary natural language queries, without manual data curation. The entire analysis—from question to integrated summary—was completed in approximately 60 seconds, compared to hours or days that would be required for manual cross-modal data integration using conventional approaches. **This represents the first system capable of translating open-ended biological questions into automated, multi-step cross-modal retrieval workflows that span molecular, morphological, and connectivity data, delivering integrated results with full provenance for reproducibility and validation.**

**Automated whole-brain tri-modal fingerprinting systematically reveals regions with cross-modal concordance versus divergence**

****

**Figure 4 |** *Automated whole-brain tri-modal fingerprinting systematically reveals regions with cross-modal concordance versus divergence.* (A) Whole-brain pairwise similarity matrices for morphology (left), molecular (center), and projection (right) fingerprints across 30 brain regions. Morphology and projection fingerprints display prominent block structures (red-orange blocks), indicating functional modularity where certain region groups share similar cellular morphologies or projection patterns. In striking contrast, the molecular fingerprint matrix is dominated by low inter-region similarity (blue), except along the diagonal, revealing that molecular composition is highly region-specific. (B) Cross-modal mismatch indices quantify the divergence between modalities. Left: Molecule-Morphology Divergence matrix shows region pairs where molecular composition diverges from morphological organization. Right: Molecule-Projection Divergence matrix reveals pairs where molecular similarity fails to predict projection patterns. Approximately 40% of region pairs display systematic cross-modal divergence (warm colors), demonstrating that concordance across modalities is not the norm. (C) Exemplar case of Molecule-Morphology Divergence: LHA versus TU. Left: Morphology feature profiles (radar plot) show nearly identical dendritic arbor characteristics between lateral hypothalamic area (LHA) and tuberal nucleus (TU). Right: Despite morphological similarity, their top neuronal subtypes are completely distinct. LHA contains diverse neuropeptidergic and neurotransmitter-defined populations (Skor1+ glutamatergic, Foxb1+ glutamatergic, Pitx2+ glutamatergic neurons), reflecting its role as a multifunctional hypothalamic hub regulating arousal, feeding, and motivated behavior. TU, by contrast, comprises a more specialized GABAergic population involved in hedonic feeding control. (D) Exemplar case of Projection Convergence despite Molecular Divergence: MOs versus LHA. Left: Projection target profiles reveal substantial overlap in target structures, with both secondary motor cortex (MOs) and lateral hypothalamic area (LHA) projecting to striatum (STR, CP, ACB), thalamus (MD, VM, ZI), and brainstem (PAG, MRN, MB). While the relative strengths differ—MOs projects more strongly to cortical and striatal motor areas while LHA emphasizes brainstem arousal centers—the shared target set reflects convergent control architecture where multiple brain systems coordinate behavior through common downstream effectors. Right: Despite this projection convergence, their molecular compositions are completely distinct. MOs comprises laminar-organized cortical glutamatergic subtypes (IT-type, ET-type, CT-type neurons marked by layer-specific genes) along with GABAergic interneurons. LHA contains hypothalamic glutamatergic neurons marked by region-specific transcription factors (Skor1+, Foxb1+, Pitx2+) and neuropeptidergic signatures.

Having established automated cross-modal integration for targeted queries, we next asked whether our system could perform systematic, whole-brain analyses of cross-modal relationships. We instructed AIPOM-CoT to construct tri-modal fingerprints (molecular, morphological, and projection) for all 32 major brain regions in our knowledge graph and compute pairwise similarity matrices for each modality (Figure 4A).

The resulting similarity patterns revealed a striking contrast. Morphology and projection fingerprints displayed prominent block structures, with certain region groups showing high mutual similarity (e.g., cortical areas in morphology, hypothalamic subregions in projection patterns). This modular organization suggests that functional relatedness at morphological or connectivity levels often transcends strict anatomical boundaries. In stark contrast, the molecular fingerprint matrix was dominated by sparse, low inter-region similarity, with high values confined almost exclusively to the diagonal. This fundamental asymmetry—modular organization at morphological and projection levels versus regional specificity at the molecular level—indicates that molecular composition is largely region-specific, whereas morphological and projection properties can converge across anatomically distant regions.

To quantify how often regions display concordance versus divergence across modalities, we computed cross-modal mismatch indices for all region pairs (Figure 4B). The Molecule-Morphology Divergence matrix identified pairs where molecular composition diverges from morphological organization, while the Molecule-Projection Divergence matrix highlighted cases where molecular profiles do not align with projection patterns. Approximately 40% of region pairs exhibited systematic cross-modal divergence, demonstrating that multi-modal concordance—where regions similar in one modality are also similar in another—is not the norm. Instead, brain organization operates along semi-independent axes, with morphological and projection similarities often arising independently of molecular composition.

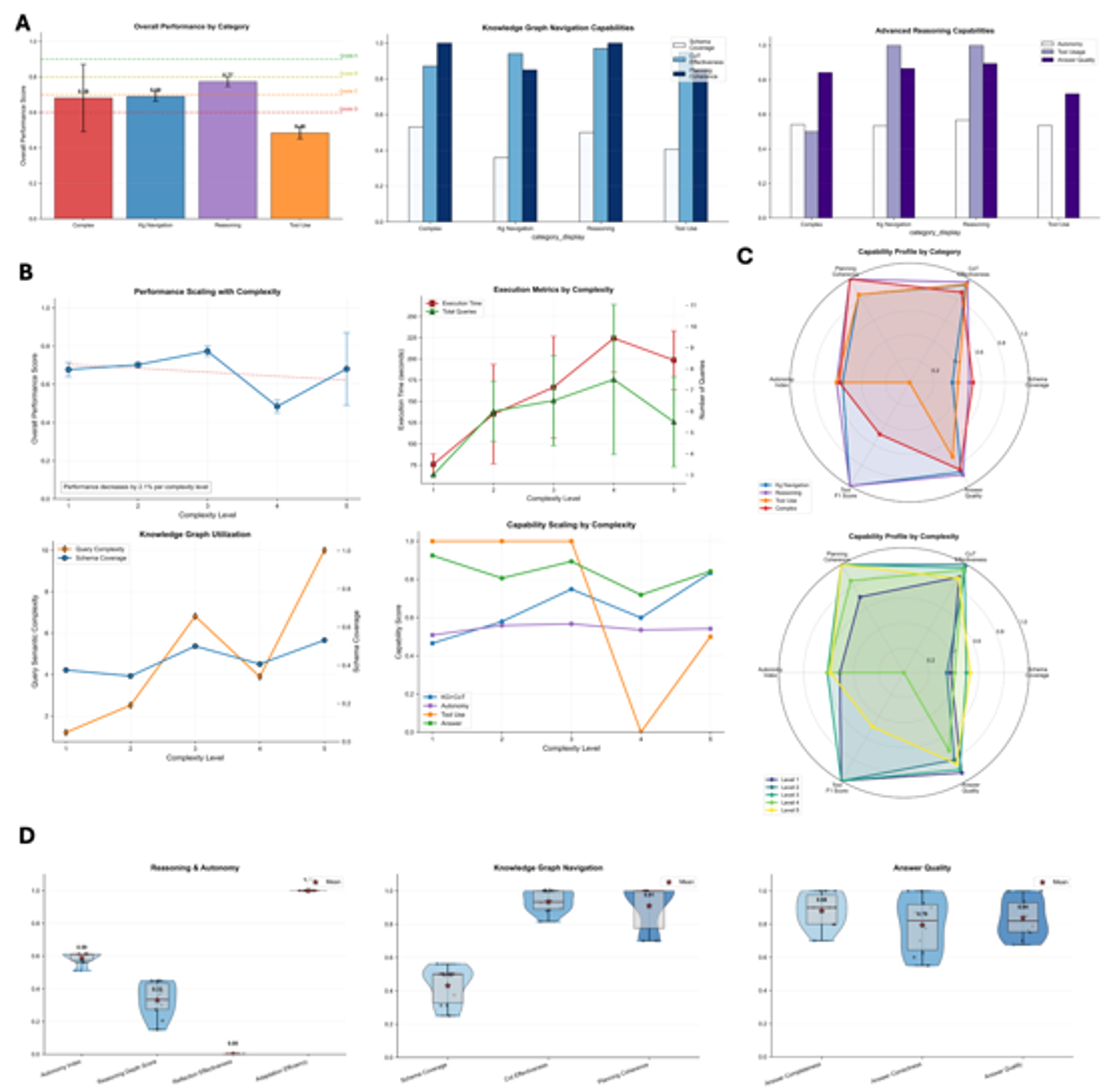
To illustrate the biological significance of these divergence patterns, we examined two representative cases automatically identified by AIPOM-CoT. First, we analyzed the lateral hypothalamic area (LHA) and tuberal nucleus (TU), which display high morphological similarity but marked molecular divergence (Figure 4C). Their morphological feature profiles—including axon length, branching complexity, and dendritic arbor characteristics—are nearly identical, suggesting similar local circuit integration strategies. However, their neuronal compositions differ dramatically. LHA contains a highly heterogeneous population including Skor1+, Foxb1+, and Pitx2+ glutamatergic neurons, along with specialized neuropeptidergic subtypes (e.g., orexin, melanin-concentrating hormone) that support its multifunctional role as a hypothalamic hub regulating arousal, feeding, reward, and motivated behavior. TU, by contrast, is dominated by a more homogeneous GABAergic population specialized for hedonic feeding regulation, receiving specific neurotensin inputs from the lateral septum.

This example demonstrates functional convergence at the morphological level despite molecular divergence. The morphological similarity likely reflects shared computational demands—both regions process feeding-related signals and may adopt similar dendritic architectures for local integration. Yet their molecular identities diverge, reflecting distinct developmental origins (LHA neurons arise from multiple progenitor domains; TU neurons derive from ventromedial hypothalamic progenitors) and functional specializations (LHA as a multifunctional integration hub versus TU as a specialized circuit node). Thus, morphological organization can converge across regions to support related functions while molecular composition remains tied to developmental history and cell-type-specific roles.

The second case examined secondary motor cortex (MOs) versus LHA, which display projection convergence despite complete molecular divergence (Figure 4D). Despite entirely different molecular compositions—MOs comprises laminar-organized cortical glutamatergic neurons (IT-type, ET-type, CT-type) while LHA contains hypothalamic glutamatergic neurons marked by Skor1, Foxb1, and Pitx2—both regions project to a largely overlapping set of targets including striatum (STR, CP, ACB), thalamus (MD, VM, ZI), and brainstem structures (PAG, MRN, MB). While the relative projection strengths differ, with MOs emphasizing cortical-striatal motor loops and LHA emphasizing hypothalamic-brainstem arousal pathways, the substantial target overlap reflects a convergent control architecture where multiple brain systems coordinate behavior through shared downstream effectors.

This convergence reveals a fundamental organizational principle: molecularly distinct neuronal populations from evolutionarily and developmentally disparate systems can converge onto common targets to enable multi-system behavioral coordination. MOs, arising from Emx1+ cortical progenitors and organized into laminar glutamatergic subtypes, provides learned motor programs and flexible sensorimotor control—answering 'how to move'. LHA, arising from hypothalamic progenitor domains and expressing region-specific transcription factors and neuropeptides, provides innate motivational drives and arousal states—answering 'why to move'. Both systems project to striatum, but with different functional contributions: MOs specifies motor sequences and habits, while LHA gates movement with motivational vigor and urgency. Similarly, both project to PAG—MOs for the motor execution of defensive behaviors, LHA for the arousal and affective components of defense.

This convergent architecture allows the brain to integrate phylogenetically ancient mechanisms (hypothalamic drives) with recently evolved capacities (cortical motor control) through shared intermediate structures. Projection patterns thus reflect functional integration demands and circuit roles rather than being determined by neurotransmitter class or molecular identity. The molecular divergence between MOs and LHA neuronal populations—reflecting distinct developmental programs, evolutionary histories, and specialized signaling repertoires—does not prevent them from accessing similar downstream targets, demonstrating that connectivity is organized by functional requirements that transcend molecular boundaries.



**Figure 5 |** Benchmarks, ablations and robustness of the executable, auditable reasoning workflow. **(**A) Overall performance and core capabilities. Left: Overall performance score by task category (mean ± 95% bootstrap CI across tasks). The score aggregates accuracy (Top K / exact match as applicable) and audit completeness (presence of effect sizes, CIs, permutation p, FDR-adjusted q, snapshot/seed and full provenance). Middle: Knowledge-graph navigation sub-capabilities (subgraph extraction, cross-context backtracking, operator-chain execution). Right: Advanced reasoning sub-capabilities (autonomous planning, reflection-driven correction, and answer interpretability). [n] tasks per category; categories defined in Methods. (B) Scaling with complexity and execution cost. Left: Overall performance as a function of task complexity level (1–5; rubric in Methods). Middle: Execution metrics—wall-clock time and total tool/LLM calls—versus complexity (mean ± 95% CI). Right: Capability scaling with complexity for planning, graph navigation, and statistical auditing. Bottom-left: Knowledge-graph utilization (graph operations per task; line) versus execution cost (bars). Bottom-right: Capability scores by complexity. (C) Capability profiles. Radar plots summarize capability distributions by category (top) and by complexity level (bottom). Values are normalized to each metric’s maximum across systems to highlight relative strengths. (D) Comparison to strong baselines. Violin/box plots of three aggregate axes—Reasoning & Autonomy, Knowledge-Graph Navigation, and Answer Quality—for our system versus baselines (graph-only pipeline, retrieval-augmented LLM, expert notebook workflow). Points are per-task scores; boxes show median and IQR; whiskers denote 1.5×IQR. Significance assessed by paired Wilcoxon tests across tasks with Benjamini–Hochberg correction; q values reported in the panel (threshold: q < 0.05). All bars/lines show mean ± 95% bootstrap CI; unit of analysis is a task instance. Exact n, effect-size definitions, permutation procedures, and multiple-testing controls are detailed in Methods.

To quantify method-level benefits beyond case studies, we assembled a task suite spanning four categories—complex queries, knowledge-graph navigation, reasoning, and free-form use—with [n] tasks per category and a five-level complexity rubric (1–5; Methods). Each task yields (i) accuracy (Top K or exact match), (ii) audit completeness (presence of effect sizes, CIs, permutation p, FDR-adjusted q, and full provenance, including snapshot and seed), and (iii) efficiency (latency, tool/LLM calls).

Across categories, the executable, auditable workflow achieved the highest composite scores (Fig. 5A, left), with the largest margins on knowledge-graph navigation tasks (Fig. 5A, middle) and consistent gains in advanced reasoning—autonomous planning, reflection, and interpretable answers (Fig. 5A, right). Improvements over baselines remained significant after multiple-testing correction (q = [ ] for category-wise paired comparisons; Methods), indicating that converting natural-language questions into operator chains with explicit statistics reliably boosts both correctness and auditable reporting.

Performance remained stable from levels 1–4 with only a modest decline at level 5 (Fig. 5B, left), while audit completeness stayed high across levels. Execution time and tool/LLM calls increased with complexity (Fig. 5B, middle) but tracked the number of graph operations invoked by the planner (Fig. 5B, bottom-left), suggesting the overhead primarily reflects intentional evidence retrieval rather than uncontrolled agent loops. Capability profiles retained their shape across levels (Fig. 5B, right; Fig. 5C, bottom), indicating that planning, graph navigation, and statistical auditing scale together rather than trading off.

Radar summaries (Fig. 5C, top) show distinct signatures: navigation-heavy tasks emphasize graph extraction/backtracking, while complex reasoning emphasizes planning + reflection + statistical auditing. Normalization across systems highlights that our approach balances these axes instead of over-optimizing a single metric.

Against a graph-only pipeline, a retrieval-augmented LLM, and an expert notebook workflow, our system shows higher medians and tighter dispersion on Reasoning & Autonomy, Knowledge-Graph Navigation, and Answer Quality (Fig. 5D). Paired Wilcoxon tests across tasks confirm significant improvements (q = [ ] for each axis). Notably, the retrieval-augmented LLM attains reasonable answer quality on low-complexity prompts but exhibits reduced audit completeness and instability at higher levels; the graph-only pipeline is precise when templates match but lacks planner flexibility.

These results demonstrate that compiling questions into auditable operator chains confers method-level advantages that persist across categories and increasing complexity, while maintaining transparent provenance and calibrated uncertainty. The benchmarking suite, statistical procedures, and minimal reproducible package (graph snapshot, task set, logs, environment, seeds) are released to support independent replication and further comparative studies.

**References**

1. Wang, Q., et al., The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. Cell, 2020. 181(4): p. 936–953.e20.

2. Oh, S.W., et al., A mesoscale connectome of the mouse brain. Nature, 2014. 508(7495): p. 207–214.

3. Ragan, T., et al., Serial two-photon tomography for automated ex vivo mouse brain imaging. Nature Methods, 2012. 9(3): p. 255–258.

4. Zhang, M., et al., A molecularly defined and spatially resolved cell atlas of the whole mouse brain. bioRxiv, 2023.

5. Yao, Z., et al., A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. Nature, 2023. 624(7991): p. 317–332.

6. Peng, H., et al., Morphological diversity of single neurons in molecularly defined cell types. Nature, 2021. 598(7879): p. 174–181.

7. Qu, L., et al., Cross-modal coherent registration of whole mouse brains. Nat Methods, 2022. 19(1): p. 111–118.

8. Liu, Y., et al., Constructing a Mouse Brain Atlas of Dendritic Microenvironments Helps Discover Hidden Associations Between Anatomical Layout, Projection Targets and Transcriptomic Profiles of Neurons. bioRxiv, 2024: p. 2024.09.22.614330.

9. Claudi, F., et al., Visualizing anatomically registered data with brainrender. Elife, 2021. 10.

10. Tyson, A.L., et al., Accurate determination of marker location within whole-brain microscopy images. Scientific Reports, 2022. 12(1): p. 867.

11. Hao, Y., et al., Integrated analysis of multimodal single-cell data. Cell, 2021. 184(13): p. 3573–3587.e29.

12. Gayoso, A., et al., Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat Methods, 2021. 18(3): p. 272–282.

13. Ashuach, T., et al., MultiVI: deep generative model for the integration of multimodal data. Nature Methods, 2023. 20(8): p. 1222–1231.

14. Yao, S., et al., ReAct: Synergizing Reasoning and Acting in Language Models. arXiv e-prints, 2022: p. arXiv:2210.03629.

15. Schick, T., et al. Toolformer: Language Models Can Teach Themselves to Use Tools. 2023. arXiv:2302.04761 DOI: 10.48550/arXiv.2302.04761.

16. Yang, H., S. Yue, and Y. He Auto-GPT for Online Decision Making: Benchmarks and Additional Opinions. 2023. arXiv:2306.02224 DOI: 10.48550/arXiv.2306.02224.

17. Jiang, S., et al., NeuroXiv: AI-powered open databasing and dynamic mining of brain-wide neuron morphometry. Nature Methods, 2025. 22(6): p. 1195–1198.

18. Dorkenwald, S., et al., FlyWire: online community for whole-brain connectomics. Nat Methods, 2022. 19(1): p. 119–128.

19. Callaway, E.M., et al., A multimodal cell census and atlas of the mammalian primary motor cortex. Nature, 2021. 598(7879): p. 86–102.

20. Lorents, A., et al., Human Brain Project Partnering Projects Meeting: Status Quo and Outlook. eNeuro, 2023. 10(9).

**Discussion**

We developed NeuroXiv-KG and AIPOM-CoT to address the fragmentation of neuroscience data across molecular, morphological, and connectivity domains. NeuroXiv-KG provides a unified substrate: a typed, cross-modal property graph encoding 34,771 nodes and 252.6 million relationships that enable direct traversal from molecular markers to anatomical regions to morphological features to projection targets. AIPOM-CoT translates natural language questions into multi-step analytical workflows with automated graph traversals, statistical computations (effect sizes, confidence intervals, permutation tests with FDR correction), and full provenance tracking (snapshot IDs, seeds, operator sequences, query hashes). Together, they transform atlas interaction from manual browsing to planned, auditable, multi-modal analyses.

The Car3+ neuron case study demonstrates automated cross-modal integration. From an open-ended prompt, AIPOM-CoT autonomously identified Car3-expressing transcriptomic subclasses, discovered CLA enrichment (~43% of Car3+ cells) through quantitative ranking, retrieved morphological reconstructions, mapped projection targets, and generated molecular fingerprints of target regions—completing in ~60 seconds an analysis requiring hours of manual effort. The whole-brain tri-modal fingerprint analysis revealed that approximately 40% of region pairs display systematic cross-modal divergence: morphological and projection patterns show functional modularity, while molecular composition is largely region-specific. Exemplar cases illustrate the mechanisms: LHA and TU share morphological features despite molecular divergence (reflecting shared computational demands versus distinct developmental origins); MOs and LHA converge on projection targets despite molecular divergence (enabling integration of cortical motor control with hypothalamic drives through shared downstream effectors). These patterns demonstrate that molecular, morphological, and connectivity features operate along semi-independent organizational axes.

Our platform occupies a distinct niche in the neuroscience ecosystem. The Allen Brain Knowledge Portal provides excellent curated resources but lacks automated cross-modal query capabilities; Blue Brain Nexus offers powerful generic infrastructure but requires users to build analytical tools; BrainGlobe excels at imaging workflows but focuses on experimental processing rather than knowledge integration; generic LLM frameworks demonstrate planning but lack neuroscience-specific structures and statistical auditing. We combine these elements: purpose-built graph design with hierarchical spatial organization and quantitative cross-modal edges, natural language interface compiled into auditable operator chains, built-in statistical discovery (correlations, permutation tests, FDR correction), and full reproducibility infrastructure. Benchmarking confirms method-level benefits: higher accuracy, audit completeness, and interpretability compared to graph-only pipelines, retrieval-augmented LLMs, or manual workflows, with advantages persisting across complexity levels.

Current limitations include region-level aggregation that blurs single-cell heterogeneity, slight stochasticity in LLM-based planning, computational cost of comprehensive permutation testing, and restriction to adult mouse brain. Future development will integrate cell-type-resolved connectivity where available, add conformal calibration for uncertainty quantification, formalize representative exemplar selection via submodular objectives, expand expert-graded task suites for benchmarking, extend the operator library to include causal inference primitives (mediation, ablation simulation), and potentially integrate with experimental platforms to enable hypothesis testing suggestions.

The pervasive cross-modal divergence has implications for conceptualizing brain organization. Molecular cell-type taxonomies, while valuable, do not fully predict morphological or connectivity features; conversely, connectivity diagrams that ignore molecular heterogeneity risk oversimplifying functional diversity. Complete understanding requires explicit multi-modal integration, treating molecular, morphological, and connectivity data as complementary rather than redundant information sources. Knowledge graphs provide the substrate for such integration, and automated reasoning agents provide the analytical machinery to explore cross-modal relationships systematically.

In summary, NeuroXiv-KG and AIPOM-CoT establish a unified platform for automated, auditable cross-modal neuroscience. By combining comprehensive data integration, natural language interfaces, built-in statistical discovery, and full provenance tracking, the system enables researchers to move from manual data browsing to planned analytical workflows that adapt to question complexity while maintaining reproducibility. This provides a foundation for systematic exploration of neuroscience’s increasingly rich multi-modal datasets.

**Methods**

**Data unification and atlas normalization**

All datasets were registered into a common coordinate system spanning three atlases: CCFv3 (population reference), CCF-ME (morphology-informed micro-environments). mBrainAligner’s coherent landmark mapping (CLM) pipeline performs cross-modal whole-brain registration and maps reconstructed neurites/somas to target space; its workflow and accuracy characterization follow Qu et al. (CLM detection-mapping; robust global + local alignment; support for fMOST/LSFM/VISoR/MRI). ME subparcellations, levels, and CP-specific examples are as in the CCF-ME atlas materials and related figures (levels, subregions, upstream/downstream examples).

**Knowledge-graph construction**

We constructed NeuroXiv-KG as a property graph with the following node types: Region, Subregion, ME-Subregion, Neuron, Transcriptomic Class/Subclass/Supertype/Cluster. Typed edges include LOCATE\_AT / LOCATE\_AT\_SUBREGION / LOCATE\_AT\_ME\_SUBREGION, PROJECT\_TO (with normalized weights), HAS\_CLASS / HAS\_SUBCLASS / HAS\_SUPERTYPE / HAS\_CLUSTER, BELONG\_TO, and morphology adjacency (DEN\_NEIGHBOURING / AXON\_NEIGHBOURING). The current build comprises 34,771 nodes and 252.6 M edges.

Region-level morphology summaries (e.g., axon/dendrite length, branch counts, max branch order) were computed per neuron and aggregated to regions/subregions; projection edges were harmonized and attached to (source region → target subregion) pairs; transcriptomic composition was derived as percentages across Class/Subclass/Supertype/Cluster tiers for MERFISH-mapped cells. (Schema and aggregation overview in Fig. 1B–D.)

**AIPOM-CoT: planning, execution, and provenance**

**Loop**

For a free-text prompt, the agent runs Think → Act → Observe → Reflect:

• Think. Parse intent; inspect the live schema; compose a task graph (e.g., select ROI → profile pattern → expand context).

• Act. Bind operators and compile to graph queries with explicit parameters and compute budgets (time, memory, max steps).

• Observe. Append estimates to an evidence buffer: effect sizes, 95% CIs, permutation-based p, FDR-adjusted q, thresholds, and sample sizes.

• Reflect. Apply policy rules: add covariates, pool tiers on sparsity, sweep thresholds on brittleness, switch metrics on instability, or Think-Deeper to expand context. Halt when coverage/stability/consistency criteria are met or the budget is reached.

**Operator library**

FIND/RANK (by property/composition/weight); PATH (constrained traversals); AGGREGATE (region/layer summaries); ENRICH (hypergeometric with FDR); SIMILARITY (cosine/JS/Euclidean); CORRELATE (Spearman; partial via residualization on covariates such as macro-division or subclass fractions); CLUSTER (hierarchical, standard linkages); VISUALIZE (ranked lists, heat maps, radar plots). All operators return results and metadata: (n), thresholds, seeds, snapshot IDs, and query hashes.

**Provenance**

Each run outputs (1) a human-readable answer with figures/tables and (2) a machine-readable trace enumerating operator, parameters, inputs/outputs, snapshot/seed, and the plan hash for exact replay.

**Tri-modal fingerprints**

For each region r:

• Molecule fingerprint: normalized composition over transcriptomic tiers (Class/Subclass/Supertype/Cluster; stacked or projected to a common K\_G-dimensional basis).

• Morphology fingerprint : region-level means (or medians) of standardized per-neuron features (axon length, dendrite length, axon/dendrite branch counts, maximum branch order, etc.), z-scored within macro-divisions to reduce cross-division scale artifacts.

• Projection fingerprint : normalized vector of outgoing PROJECT\_TO weights over target subregions (or atlas-consistent targets).

This fingerprinting is the basis for region×region similarity maps used in Fig. 4.

**Similarities, distances, and mismatch indices**

Within each modality, we compute pairwise similarities S and derive distances D=1-S (for cosine-based cases) as follows—matching what is described/used in the results:

•Molecular (composition):

•Morphology (standardized features):

Primary: cosine similarity on z-scored feature vectors

(Euclidean on z-scores was used as a robustness check; conclusions were unchanged.)

• Projection (normalized target distributions):

From these, we define two cross-modal mismatch indices (higher = stronger disagreement):

These indices generate the “molecular-vs-morphology” and “molecular-vs-projection” maps in Fig. 4 and are used to rank exemplar pairs for evidence panels.

**Live-query exemplar (Car3)**

The agent decomposes “tell me about Car3⁺ neurons” into region enrichment, neuron retrieval, target profiling, and composition analysis, focusing on CLA as the top-ranked region and surfacing repeatedly hit targets (ENTl, MOs/ACAd, CP/SNr) with their molecular makeups; the full run is exported as a provenance subgraph and execution trace.

**Data availability**

**Code availability**

**Supplementary Notes**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Framework** | **Year** | **Domain** | **Statistical Rigor** | **Schema Adaptation** | **Provenance** | **Our Advantage** |
| ReAct | 2022 | General QA | ❌ | ❌ | Trace only | ✅ Statistical auditing |
| Toolformer | 2023 | General NLP | ❌ | ❌ | ❌ | ✅ Multi-round reflection |
| AutoGPT | 2023 | Task planning | ❌ | ❌ | ❌ | ✅ Controlled execution |
| LangChain | 2023 | General | ❌ | ❌ | Partial | ✅ Domain KG integration |
| LangGraph | 2024 | General | ❌ | ❌ | Full | ✅ Neuroscience-specific |
| **AIPOM-CoT** | 2025 | Neuroscience KG CoT | ✅ Permutation tests | ✅ Cypher generation | ✅ Snapshot + seed | **All combined** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dimension** | **Allen BKP / CTKE** | **Blue Brain Nexus** | **BrainGlobe suite (cellfinder/brainreg, etc.)** | **NeuroXiv 2.0 (NeuroXiv-KG + AIPOM-CoT)** |
| **Data integration & resolution** | ✅ Curates multi-modal **cell-type** data (transcriptomics, morphology, e-phys) with strong browsers; resolution: **single-cell**; but **whole-brain morphology→projection unification** is limited (❌ comprehensive end-to-end unification). | ✅ Generic, powerful **KG/data management** backbone for heterogeneous neuro data and multi-scale entities; more of a **substrate** than a ready unified analysis stack (❌ out-of-the-box unified analysis). | ✅ Strong for **3D imaging pipelines** (cell detection, registration, atlas mapping); resolution driven by images; essentially **not** a multi-modal KG (❌ multi-modal KG). | ✅ Whole-brain, **single-neuron** **CCFv3 standardization**; **four-level spatial hierarchy** + **four-level cell hierarchy** designed to add/connect molecular/projection layers **inside one graph**. |
| **Graph structure & cross-modal links** | ✅ Anchored in cell-type ontology; exploratory, **flexible cross-modal edges** are relatively limited (✅ *partial* / ❌ limited flexibility). | ✅ True KG (RDF/ontologies) with rich entity/relationship modeling; users define schemas and queries. | ❌ No unified KG; coordination primarily via atlas coordinates and scripts. | ✅ Neuroscience-specific **purpose-built KG**: hierarchical regions + cell ontology + neuron nodes + cross-modal edges (morphology, projections, statistics); **de-duplicated adjacency** and **uniform projection schema** for robust graph algorithms. |
| **Query & interaction** | ✅ Excellent **portal UI** and structured browsing; ❌ no natural-language (NL) or user-defined ad-hoc graph pattern queries. | ✅ Structured access (REST/SPARQL); power users write queries; ❌ limited exploratory UX out-of-the-box. | ✅ Script/GUI-driven; users compose pipelines; ❌ no NL or general graph-query layer. | ✅ **Dual mode: GUI + NL dialog**. NL requests are compiled into multi-step graph/analysis plans; supports iterative, exploratory Q&A over the KG. |
| **Statistics / causal / pattern discovery** | ✅ Precomputed summaries/visualizations; ❌ no on-platform ad-hoc statistics or causal tools. | ✅ Enables discovery via data access; analysis lives in user tools (❌ no built-in statistics/causal). | ✅ Produces quantitative outputs (counts/coordinates) for **downstream** stats; ❌ not a statistics engine / no causal. | ✅ Built-in **pattern discovery** (correlations, partial-ρ, permutation + FDR, effect sizes); **non-paired** cross-modal analysis via region-level aggregation; designed to extend to mediation/ablation **“WHY”** operators. |
| **AI agent / reasoning** | ❌ | ❌ | ❌ | ✅ **AIPOM-CoT**: schema-adaptive **Chain-of-Thought** agent that plans → executes → observes → reflects; auto-generates queries/analyses with confidence calibration. |
| **Workflow replay / reproducibility** | ✅ Data are citable; ❌ portal does **not** version user analysis steps. | ✅ Strong data/ontology **provenance & versioning**; ✅ (*partial*) analysis reproducibility is user-built. | ✅ Repro depends on user scripts/env; ❌ no unified workflow provenance. | ✅ **Snapshot + query hash + seeds + machine-readable traces**; portal re-runs are **bit-for-bit** identical; expanding to full workflow versioning. |
| **Biological discovery output** | ✅ **Indirect** (platform curates datasets that support discoveries). | ✅ **Indirect** (platform enables projects that produce discoveries). | ✅ **Indirect** (tools feed studies; platform itself doesn’t mine). | ✅ **Direct discovery support**: tri-modal fingerprints, whole-brain similarity and **mismatch maps**; case studies (e.g., CLA/Car3) linking quantitative patterns → testable circuit hypotheses. |
| **“Question → Insight” closed loop** | ✅ High-quality browsing; ❌ limited analytical closure. | ✅ Strong data backbone; ❌ exploratory closure is external. | ✅ Strong imaging analytics; ❌ limited knowledge/insight loop. | ✅ **End-to-end loop**: NL question → KG traversal → statistical audit → confidence scoring → **replayable** report. |

**Acknowledgements**

This work was mainly supported by several initiatives of neuroscience awarded to Hanchuan Peng.

**Author contributions**

**Competing interests**

The authors declare no competing interests.