

From Public Repositories to Target Hypotheses An End-to-End Data-to-Insights Journey for scRNA and Spatial Omics with Knowledge Graphs

Ardigen: Ida Kupś, Sergiusz Wesołowski, Jakub Widawski, Michał Stachowiak, Marzena Mura, Magdalena Ochab, Patrycja Marciniak, Krzysztof Kolmus, Marek Kudła, Przemysław Kapusta, Błażej Szczerba, Dawid Rymarczyk, Jan Majta

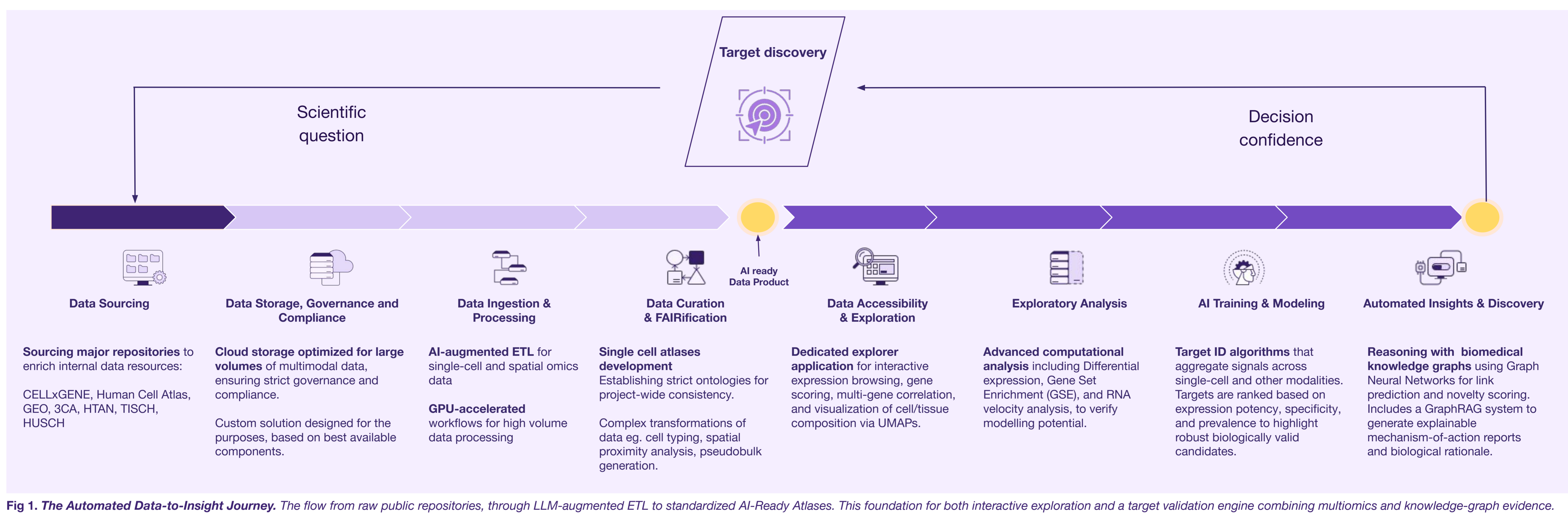


Fig 1. The Automated Data-to-Insight Journey. The flow from raw public repositories, through LLM-augmented ETL to standardized AI-Ready Atlases. This foundation for both interactive exploration and a target validation engine combining multiomics and knowledge-graph evidence.

ABSTRACT

Single-cell and spatial datasets hold immense potential but are hindered by curation bottlenecks and incompatible annotations. We present a scalable "Data-to-Insights Journey" that converts these sources into actionable target hypotheses through three core pillars:

- 1) LLM-augmented ETL, combined with atlas generation workflow, reduces processing time from days to minutes, retaining 90% accuracy. It addresses data fragmentation by unifying metadata and establishing common standards.
- 2) A scientist-ready browser for immediate access to AI-ready data products, for biologists to visually explore and validate datasets without coding.
- 3) Multiomics target ranking algorithm drives discovery by integrating single-cell data with other modalities. It synthesizes this multimodal landscape into prioritized targets. Furthermore refines the ranking with knowledge-based second line of evidence.

This end-to-end workflow accelerates the transition from raw data to reproducible insights, and has resulted in a discovery of a novel target that was later validated.

BACKGROUND

The exponential growth of public single-cell and spatial transcriptomics repositories offers an unprecedented resource for precision medicine and target discovery. However, the sheer volume and heterogeneity of this data create a paradox: while information is abundant, actionable insight remains scarce. The primary bottleneck lies in the lack of standardization; datasets are often siloed by inconsistent metadata, varying normalization methods, and assay-specific biases.

Currently, researchers spend a disproportionate amount of time, often up to 60%, on manual data wrangling. Essential tasks such as harmonizing annotations, mapping ontologies, and correcting for batch effects are labor-intensive and error-prone. Furthermore, the complexity of spatial omics, exacerbates these challenges.

To transition from ad-hoc analysis to systematic discovery, there is a critical need for automated, scalable infrastructure. We must move beyond simple data storage to create "AI-ready" ecosystems where raw public data is rigorously curated, harmonized, and transformed into high-fidelity products capable of powering robust downstream modeling, multimodal integration, and target hypothesis generation.

IMPACT

- Reduces manual curation burden from days to under 30 minutes.
- Facilitates immediate, code-free visual interrogation of complex spatial atlases.
- Prioritizes high-confidence, validated, targets via synergistic multi-omics and knowledge-based evidence.

METHODS / TECHNOLOGIES

- 1. Sourcing & Cloud Infrastructure:** Aggregation of diverse datasets including GEO, CELLXGENE, Human Cell Atlas, HTAN, and 3CA. These are stored on a scalable, cloud-native platform optimized for managing massive, multimodal volumes, serving as the foundation for our high-throughput pipeline.
- 2. AI-Augmented Ingestion & Processing:** To overcome the bottleneck of manual curation, we deploy LLM-augmented data engineering tools. These models automate metadata parsing and harmonization at the ingestion stage, ensuring consistency across heterogeneous sources. Heavy data processing tasks are accelerated via GPU infrastructure based on NVIDIA RAPIDS workflow, enabling the rapid transformation of raw inputs into structured formats.
- 3. Creation of AI-Ready Data Products (Atlases):** We compile high-fidelity Atlases through a rigorous standardization process. This involves:
 - **Ontology definition:** Establishing controlled vocabularies and coherent terminology (e.g., Ontocree) to align disjointed datasets.
 - **Advanced annotation:** Performing consistent cell type annotation and, for spatial omics, integrating proximity analysis to capture tissue architecture.
 - **Assay-specific optimization:** Utilizing assay-specific tools to mitigate technical noise before integration.
- 4. Scientist-Centric Exploration & EDA:** We provide an intuitive Spatial & scRNA Data Browser that allows researchers to visually inspect curated Atlases. This tool facilitates Exploratory Data Analysis (EDA), allowing users to perform "ROI checks" on data feasibility before committing to heavy computation. The environment seamlessly supports pre-existing Python and R-based workflows.
- 5. Multimodal Target Ranking:** For hypothesis generation, we deploy proprietary Target ID algorithms. These utilize multimodal data products, including multi-level pseudobulking to aggregate signals across single-cell and spatial modalities. The system integrates these layers to rank targets based on expression potency, specificity, and prevalence, filtering noise to highlight robust biological candidates.
- 6. Knowledge-Graph Augmented Refinement:** To validate data-driven insights, we integrate a Biomedical Knowledge Graph (leveraging PrimeKG and literature) centered on relevant biology (e.g., immune pathways)
 - **Link prediction:** We employ Graph Neural Networks (GNNs) to predict novel "drug-treats-disease" links.
 - **Novelty scoring:** Algorithms like Personalized PageRank filter for high-impact, novel targets.
 - **Reliability:** A GraphRAG system generates mechanism-of-action reports, bridging the gap between statistical ranking and biological rationale.
 - **Automation:** The system works in a loop upon new data ingestion, refining the target hypothesis list and updating the findings.

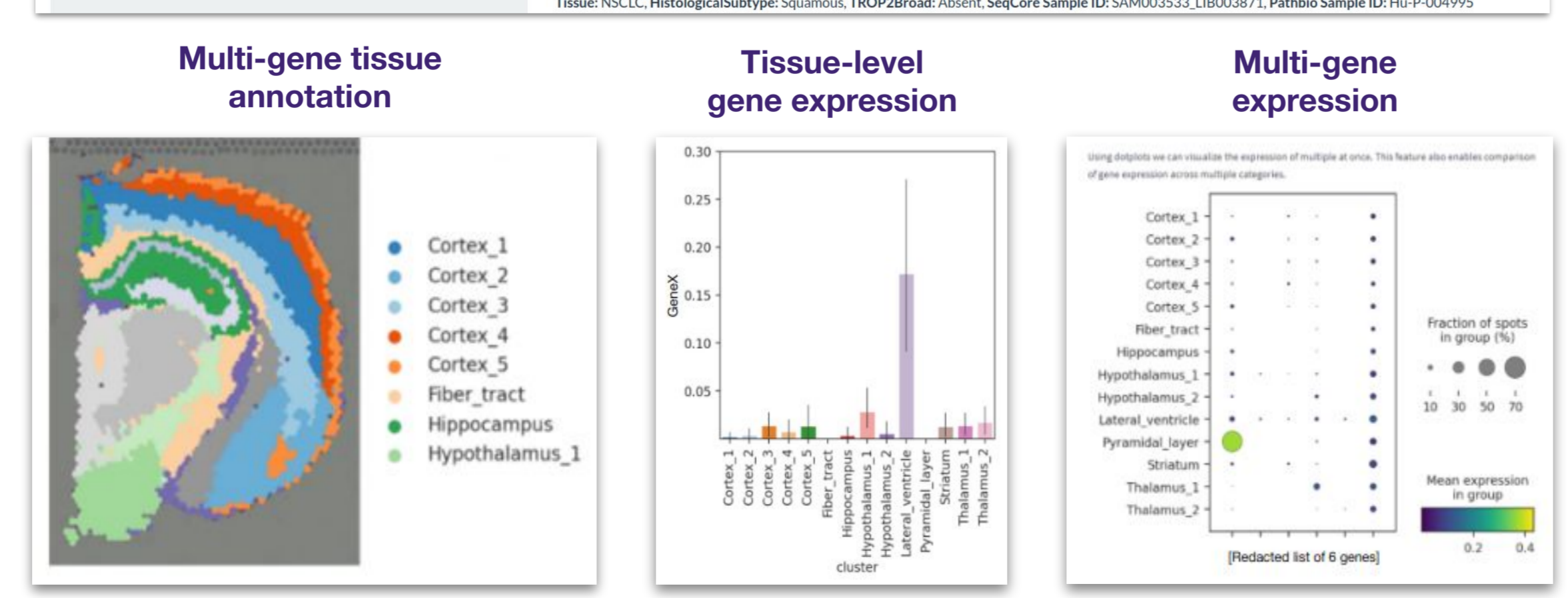
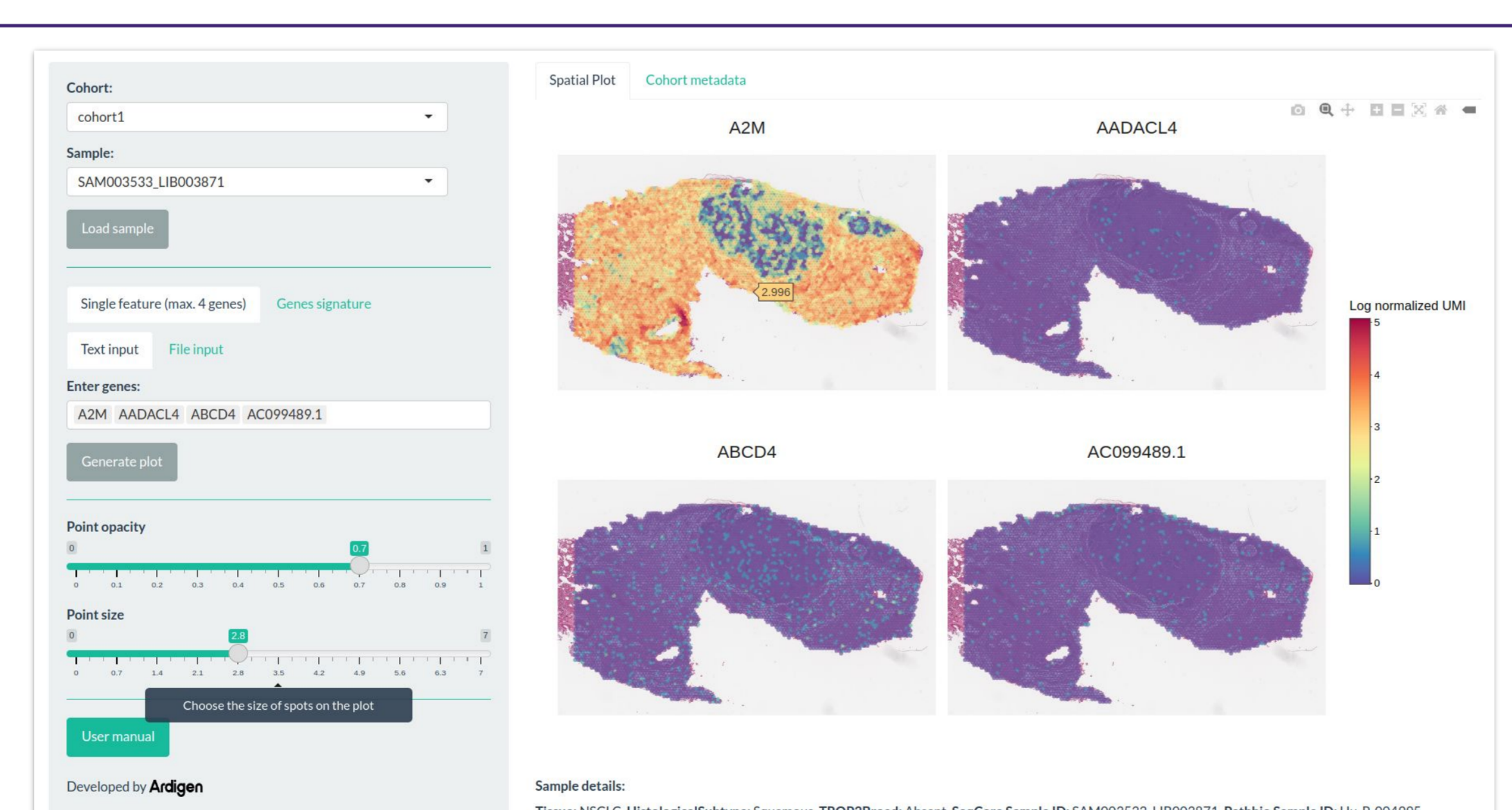


Figure 2. Scientist-Centric Spatial Data Browser. An interactive, code-free interface for exploring AI-ready atlases. It combines real-time gene signature visualization on tissue architecture (top) with automated modules for expression profiling and cluster comparisons (bottom).

LLM Cleanup	Justifications	Condition
None		Hippocampus, PSEN1-E280A + APOEε4 carrier
Alzheimer's disease		The document describes a patient with the PSEN1 E280A mutation and APOE3 Christchurch variant, specifically referencing autosomal dominant Alzheimer's disease and related tau pathology, indicating the sample disease is Alzheimer's disease.

Condition	Tissue	Cell type	Series title	Series overall design	Cells
None	Brain	Frontal cortex	Distinct tau neuropathology and cellular profiles of a APOE3 Christchurch homozygote protected against Autosomal Dominant Alzheimer's dementia	We performed comparative gene expression profiling analysis using data obtained from three different brain regions (Frontal cortex, hippocampus and occipital cortex) of a post mortem brain of a patient carrier of the autosomal dominant PSEN1-E280A mutation and homozygous for the APOE3 Christchurch variant.	Exc 2
Alzheimer's disease	Brain	excitatory neuron			

Fig 3. Automated Metadata Harmonization with LLM Reasoning. A comparison table contrasts the original metadata (top row) against the LLM-standardized output (bottom row), transforming unstructured inputs into high-quality annotations. "Justification" helper table provides the evidence-based reasoning used by the model to validate the classification.

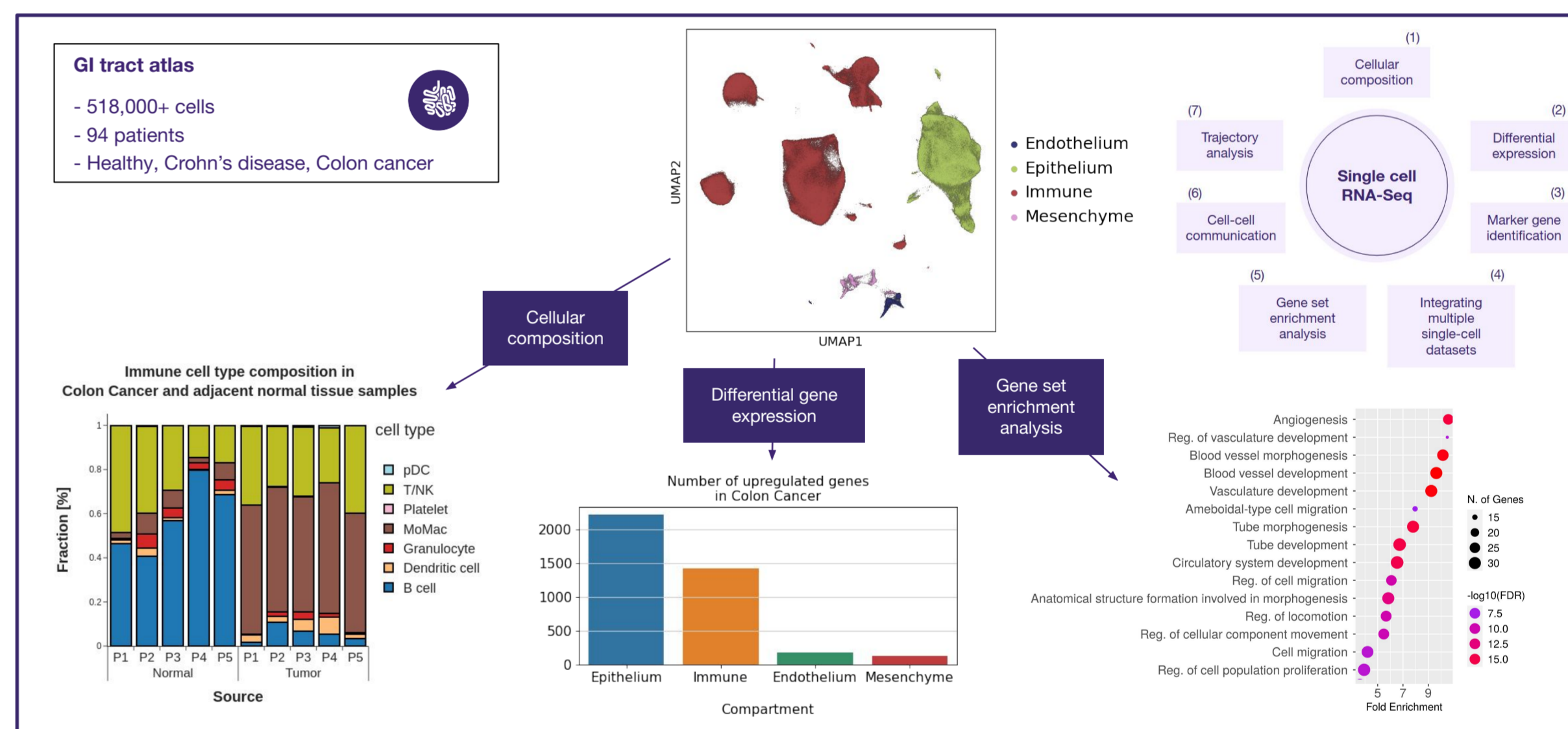


Fig 4. Single-cell and spatial transcriptomics analysis. The platform utilizes cellular composition, differential expression, and gene set enrichment across cells from healthy and disease states, enabling deep insight into tissue architecture and disease-related molecular changes.

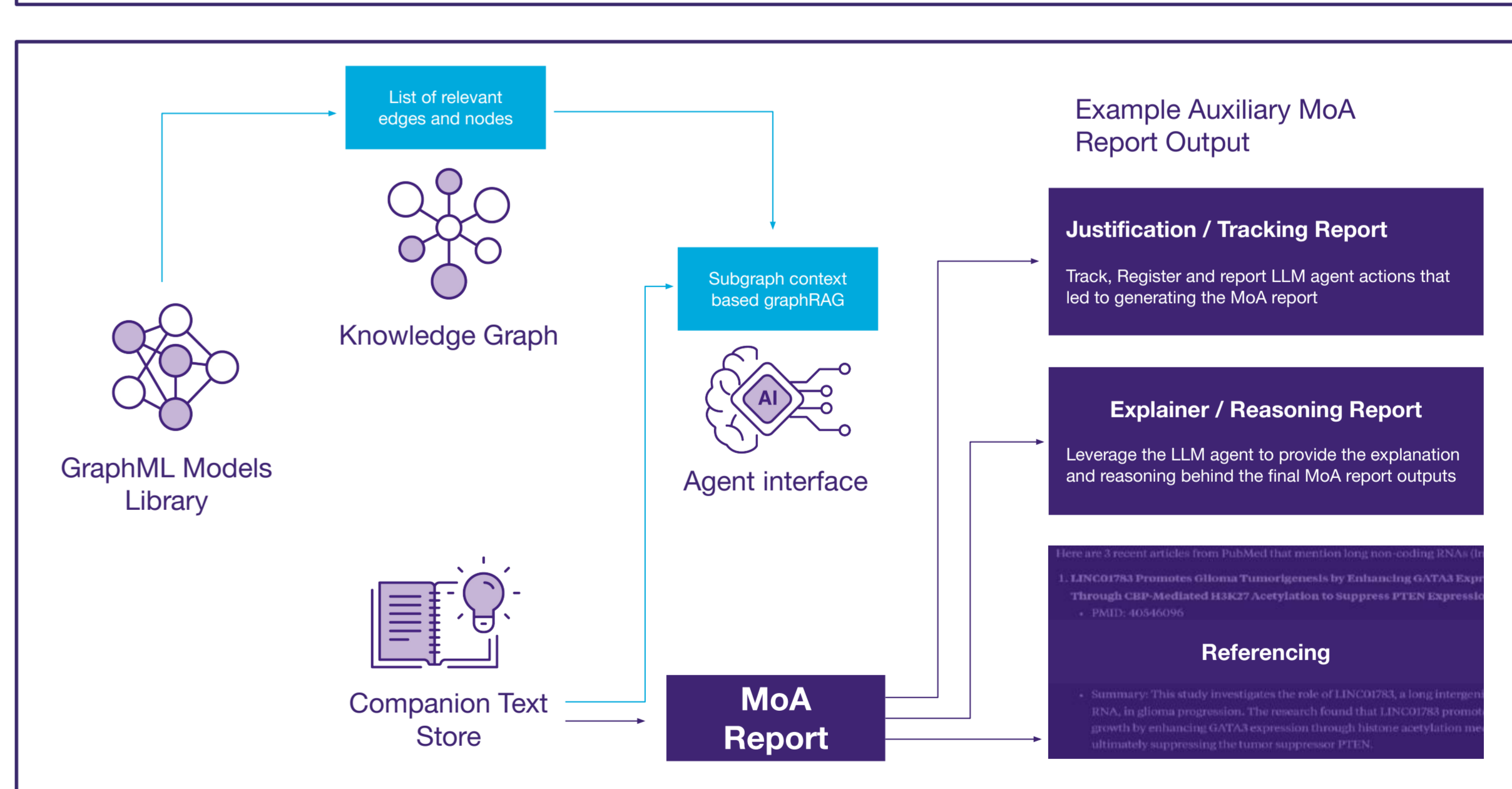


Fig 5. AI-Augmented MoA Discovery. The engine integrates GraphML with a Knowledge Graph for validation. An AI Agent uses GraphRAG to synthesize detailed MoA reports, delivering transparent, evidence-based reasoning via justification and referencing layers.

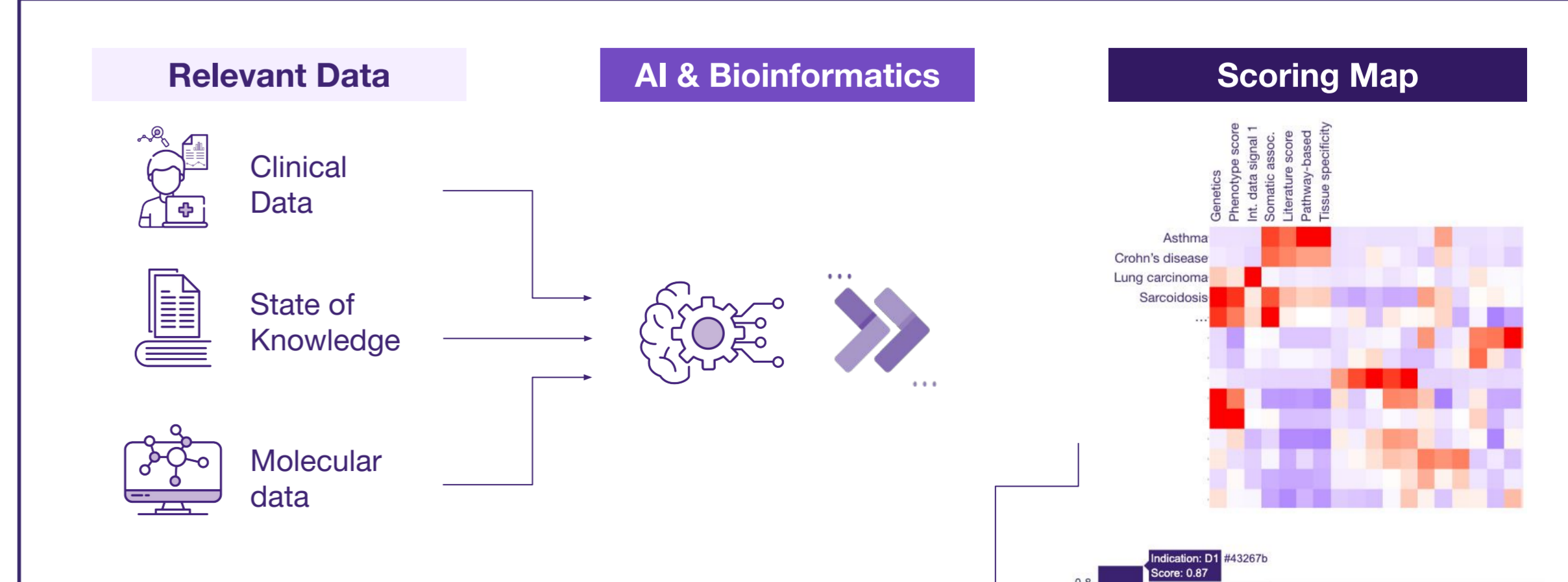


Fig 6. AI-driven target prioritization for unbiased discovery. Relevant data are integrated using AI and bioinformatics methods to generate prioritization scores, enabling transparent final ranking of drug targets across multiple disease indications.

REFERENCES
 [1] Barrett T et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013 Jan;41(Database issue):D991-5. doi: 10.1093/nar/gks1193. Epub 2012 Nov 27. PMID: 2318256; PMCID: PMC3531084.
 [2] Ochoa D et al. Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. Nat Rev Drug Discov. 2022 Aug;21(8):551. doi: 10.1038/s41573-022-00120-3. PMID: 35804044.
 [3] Minkiel EV et al. Refining the impact of genetic evidence on clinical success. Nature. 2024 May;629(8012):624-629. doi: 10.1038/s41586-024-07316-0. Epub 2024 Apr 17. PMID: 39632401; PMCID: PMC11096124.
 [4] Liu C et al. A probabilistic knowledge graph for target identification. PLoS Comput Biol. 2020 Apr 5;20(4):e1011945. doi: 10.1371/journal.pcbi.1011945. PMID: 32573905; PMCID: PMC7110345.
 [5] Markowska M et al. Synthetic lethality prediction in DNA damage repair, chromatin remodeling and the cell cycle using multi-omics data from cell lines and patients. Sci Rep. 2023 Apr 29;13(1):7049. doi: 10.1038/s41598-023-34161-4.
 [6] Sayers EW et al. Database resources of the National Center for Biotechnology Information in 2025. Nucleic Acids Res. 2025 Jan 6;53(D1):D20-D29. doi: 10.1093/nar/gkae979. PMID: 39526373; PMCID: PMC11701734.