A CONVERSATIONAL MULTI-AGENT AI FRAMEWORK FOR INTEGRATED MULTI-OMICS ANALYSIS AND BIOMEDICAL DISCOVERY

Pankaj Rajdeo[1,†], **Shunya Asanuma**[1], **Michal Kouril**[1], **Peixin Lu**[1], **Jichao Chen**[1]
Aman Chadha[2], **V. B. Surya Prasath**[1], **Bruce J. Aronow**[1,†], **Nathan Salomonis**[1,†]

[1]Cincinnati Children's Hospital Medical Center, USA
[2]Google DeepMind, USA

[†]Corresponding authors.
{pankaj.rajdeo,bruce.aronow,nathan.salomonis}@cchmc.org

ABSTRACT

Single-cell and spatial transcriptomics offer unprecedented opportunities to decipher the mechanisms of disease and tissue development, however, this process requires teams of experts, iterative trial-and-error and reasoning across modalities. Here we present a conversational agentic framework for integrated multi-omics reasoning and biomedical discovery. It is deployed as a hierarchical multi-agent architecture in which a stateful supervisor decomposes natural-language research questions into parallel, tool-grounded tasks executed by specialized agents for single-cell analysis, spatial transcriptomics, literature synthesis, drug repurposing, and functional enrichment. A virtual filesystem enables efficient inter-agent communication while preventing context degradation, and the framework maintains long-term memory for personalized analytical context across iterative sessions. For therapeutic hypothesis generation, the framework includes Direction-Aware Repurposing and Targeting (DART), a method that distinguishes perturbations that reverse disease transcriptional programs from those that reinforce or disrupt them, enabling cell-type–resolved therapeutic prioritization and safety profiling. Applied to 39 datasets, spanning fetal development, tissue homeostasis and 15 pathological conditions, the system generates publication-ready results with complete parameter provenance for reproducibility. We demonstrate autonomous discovery that identifies known drivers of different lung diseases, cell-communication networks, and validated FDA-approved drugs. Notably, DART predicts Saracatinib (SRC inhibitor) as a top therapeutic candidate, now in Phase 1b/2a trials (NCT04598919) while identifying cell-type-specific safety risks and enables efficacy and safety profiling at single-cell resolution. This architecture is tissue-agnostic, offering a blueprint for agentic AI systems that integrate foundation models with executable, verifiable scientific discovery. We instantiate this framework as LungChat, deployed for lung biology by the NHLBI LungMAP Consortium (https://chat.lungmap.net).

1 INTRODUCTION

Advances in single-cell and spatial transcriptomics have fundamentally changed how biological systems are studied, enabling cell-type–resolved and spatially contextualized analyses across development, homeostasis, and disease. Large-scale atlas efforts such as The Human BioMolecular Atlas Program (HuBMAP) Börner et al. (2025), Human Cell Atlas (HCA) Regev et al. (2017); Rozenblatt-Rosen et al. (2017), and LungMAP Gaddis et al. (2024) have generated millions of profiles across the spectrum of human tissues, development and disease, creating unprecedented opportunities for systems-level discovery. However, the increased scale and complexity of these datasets, spanning dozens to hundreds of conditions, shifts the challenge from analysis to systems-level integration. Extracting biological insight increasingly requires integrating heterogeneous data modalities, recon-

ciling inconsistent metadata, selecting appropriate analysis methods, and synthesis of results from challenging multi-step workflows.

Existing solutions address parts of this problem but remain structurally limited. Interactive visualization platforms such as the CZI CellxGene, and UCSC cell browsers Program et al. (2025); Speir et al. (2021) enable exploration of atlas-level datasets but lack support for deep, quantitative analysis. General-purpose AI assistants (e.g., ChatGPT, Claude, etc) can synthesize vast amounts of prior knowledge from the literature but lack harmonized quantitative data, metadata or access to validated analysis tools, resulting in "best guesses" Ji et al. (2023). As a result, there is a growing disconnect between the availability of atlas-scale data and the ability of researchers to interrogate it rigorously and iteratively.

Recent progress in large language models (LLMs) and agentic AI systems suggests a path forward. Conversational interfaces have demonstrated promise for lowering the barrier to complex analyses Schaefer et al. (2025), while multi-agent architectures enable decomposition of complex tasks across specialized components. However, most existing systems focus on a single data modality, emphasize autonomous hypothesis generation over direct data analysis and interpretation, or lack mechanisms for executing real workflows with reproducible outputs.

In this work, we introduce a conversational multi-agent AI framework for integrated multi-omics analysis, designed to support transparent, reproducible biomedical discovery through natural language interactions. The framework combines (i) **multi-agent orchestration**, in which a supervisor coordinates specialized analytical agents; (ii) **tool-augmented reasoning**, grounding language model outputs in structured execution of validated bioinformatics methods including functional enrichment (ToppGene Chen et al. (2009)), drug connectivity and repurposing (iLINCS Pilarczyk et al. (2022)), literature synthesis (PubMed); and (iii) **provenance-tracked reproducibility**, where every analytical step produces persistent, publication-ready outputs with complete parameter provenance.

Applied to idiopathic pulmonary fibrosis (IPF), we demonstrate that the framework independently prioritizes Saracatinib, a finding validated by its subsequent advancement to clinical trials. In Chronic Obstructive Pulmonary Disease (COPD), which affects over 300 million individuals, DART correctly identifies Fluticasone Propionate (Flonase), a drug already commonly used to improve lung function and decrease COPD-associated mortality.

We demonstrate this framework through LungChat, an open-access analysis interface for lung molecular omics integrating single-cell, spatial, and regulatory interaction data within a unified conversational interface. LungChat supports static and interactive visualizations (Morpheus, Vitessce, ShinyCell, Azimuth) and literature-grounded analysis via PubTator3 Wei et al. (2024) and EuropePMC Rosonovski et al. (2024). The full tool inventory and analytical capabilities are detailed in Section 3.3. For therapeutic hypothesis generation, we developed Direction-Aware Repurposing and Targeting (DART), a method for cell-type-resolved drug prioritization that classifies perturbations as repression or activation to predict on-target efficacy and off-target safety at single-cell resolution (Section 3.7).

By aligning natural language interaction with real analytical execution, this work positions conversational multi-agent systems as a practical interface for atlas-scale data exploration. The framework complements emerging AI scientist approaches Lu et al. (2024a); Gottweis et al. (2025) by prioritizing transparent, tool-grounded analysis over autonomous hypothesis generation, addressing a critical gap between accessibility and rigor in contemporary biomedical research.

2 Related Work

Recently described pretrained language models and single-agent approach, such as CellWhisperer Schaefer et al. (2025), CompBioAgent Zhang et al. (2025), Cell2Text Kharouiche et al. (2025) and scChat Lu et al. (2024b), provide initial conversational interfaces for single-cell genomics cell-type curation and automated data visualization. These approaches demonstrate the utility of chat-based interfaces for primary single-cell dataset analysis, but they do not address many of the fundamental challenges faced by biomedical researchers, including standardized integration of findings across studies and modalities (space, time, and disease); functional prediction (pathways, regulatory potential, and cellular interactions); assessment of the potential positive and negative impacts of new therapies; and contrasting findings across disease and developmental contexts.

Recently, multi-agent architectures have begun to emerge, with an emphasis on multimodal data analysis and reasoning. For example, BioMaster Su et al. (2025), BioAgents Mehandru et al. (2025) and Biomni Huang et al. (2025) provide programmatic logic for intra-agent communication, planning and task execution to nominate hypotheses for validation. Specialized drug-target prediction tools, such as DrugAgent Inoue et al. (2024) and TxGNN Huang et al. (2024), introduce foundational zero-shot or knowledge-graph based models for drug repositioning, with high accuracy relative to expert evaluation. While extremely valuable, these approaches do not yet leverage quantitative genomics data to generate and refine hypotheses for real or in silico evaluation.

Advanced stand-alone analytical frameworks for automated gene regulatory prediction and spatial transcriptomics analysis now provide the means to probe cell-interactions and niches at a cellular and subcellular resolution. These includes CellChat Jin et al. (2021) and Squidpy Palla et al. (2022) and Nicheformer Schaar et al. (2024) for neighborhood enrichment and cell-cell communication inference. Nonetheless, we currently lack the ability to sufficiently integrate the results of these and related approaches across spatial technologies and disease cohorts.

Our framework addresses these gaps through executable multi-omics workflows spanning single-cell, spatial, literature, and perturbation analysis with provenance-tracked reproducibility. Unlike autonomous AI scientist systems Mitchener et al. (2025); Gottweis et al. (2025) that emphasize hypothesis generation, our approach prioritizes transparent, tool-grounded execution with verifiable outputs. Table D1 (Appendix D) provides a capability-level comparison across systems.

3 METHODS

3.1 PROBLEM FORMULATION AND SCOPE

We formalize the task of interactive multi-omics analysis as follows. Given a natural language query $q \in \mathcal{Q}$, a collection of multi-omics datasets $\mathcal{D} = \{D_{sc}, D_{st}, D_{pb}\}$ spanning single-cell RNA-seq, spatial transcriptomics, and pseudobulk expression data, and a library of analytical tools $\mathcal{T} = \{t_1, \ldots, t_n\}$, our goal is to generate a structured response $\mathcal{R} = (\mathcal{V}, \mathcal{I}, \mathcal{A})$ comprising visualizations $\mathcal{V}$, biological insights $\mathcal{I}$, and reproducible outputs $\mathcal{A}$.

The central challenge lies in decomposing unstructured scientific queries into executable analytical workflows while ensuring biological accuracy, cross-modal data integration, and full reproducibility. Unlike single-turn question answering, interactive analysis requires maintaining analytical state across multi-step investigations where subsequent queries build upon prior results.

3.2 SYSTEM OVERVIEW

LungChat is an AI system for integrated multi-omics analysis, designed to support biomedical discovery through interactive analysis. The system employs a hierarchical multi-agent architecture where a supervisor agent orchestrates specialized sub-agents, each augmented with domain-specific analytical tools (Figure 1). User queries are decomposed into tractable sub-tasks that execute in parallel when independent, with results synthesized into coherent scientific responses.

The architecture integrates three core capabilities: (1) multi-agent orchestration for complex query handling, (2) tool-augmented reasoning enabling precise analytical operations, and (3) unified access to heterogeneous omics data through ontology-grounded harmonization. Intermediate analytical outcomes serve as feedback signals that update the supervisor's task decomposition, analytical resolution, and operator selection, establishing a computational lab-in-the-loop paradigm where data-driven results guide subsequent analytical actions without retraining or parameter updates.

3.3 ANALYTICAL WORKFLOWS

3.3.1 OMICS DATASETS

LungChat operates over precompiled single-cell gene expression and spatial transcriptomics omics datasets assembled from ¿ 37 studies curated by the LungMAP Consortium and the Human Cell Atlas initiative. Single-cell datasets are provided as CellxGene-curated `.h5ad` objects with consistent, standardized mappings to CellxGene-supported biomedical ontologies, including harmonized

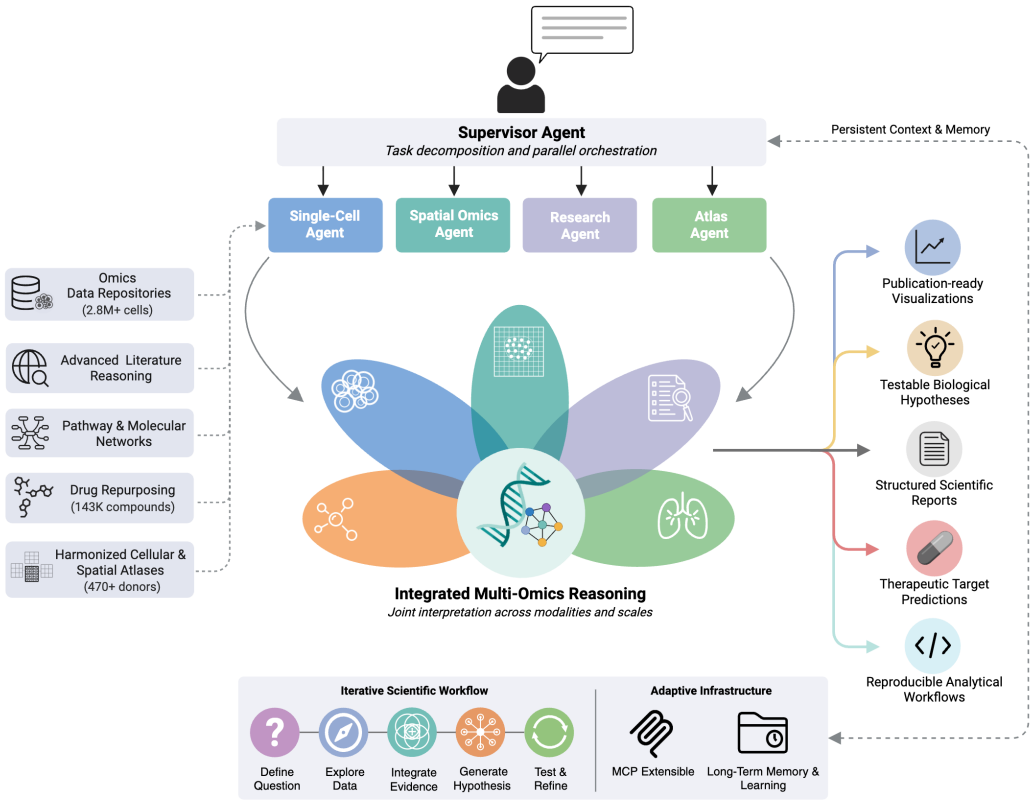

Figure 1: **LungChat system architecture and workflow.** A hierarchical multi-agent framework integrating conversational AI with multi-omics analysis. The Supervisor Agent orchestrates four specialized sub-agents (Single-Cell, Spatial, Research, Atlas) that access curated data repositories (scRNA-Seq, spatial transcriptomics, drug compounds, regulatory networks, literature) to generate publication-ready outputs including visualizations, testable hypotheses, therapeutic predictions, and reproducible workflows. The system supports iterative scientific workflows (Define → Explore → Integrate → Generate → Test & Refine) with long-term memory and MCP-extensible infrastructure for adaptive tool integration.

cell-type, disease, developmental stage, and assay annotations. To enable interactive analysis at scale, metacells and differential expression statistics are precomputed.

In addition to expression measurements, regulatory priors derived from ChIP-seq experiments and protein–protein interaction networks from prior studies are embedded locally within each `.h5ad` object using the NetPerspective software. This enables rapid retrieval of gene regulatory and interaction context during downstream analyses.

Spatial transcriptomics data include matched image-based (Xenium) and whole-transcriptome (Visium-HD) profiles. Integration between Xenium and Visium-HD datasets is performed through spot- and cell-segmented alignment and annotation stability that enables cross-platform comparison of spatial gene expression patterns and tissue organization across technologies. Appendix A provides details of the dataset curation and preprocessing undertaken.

3.3.2 Specialized Tools

Analytical workflows in LungChat are executed through a library of 51 specialized analytical and knowledge tools distributed across four specialist agents: the Single-Cell agent (13 tools for differential expression, visualization, and correlation analysis), Spatial agent (26 tools for spatial mapping,

cell-cell communication, and neighborhood enrichment), Research agent (4 tools for functional enrichment, drug connectivity, literature search, and clinical trials), and LungMAP agent (8 tools for consortium data access). Single-cell tools are implemented in Python using Scanpy, Matplotlib, Seaborn, and iGraph; spatial transcriptomics tools are implemented in R using Seurat, CellChat, and Squidpy. Table B1 (Appendix B) provides a complete reference for analytical, external API, and knowledge tools.

All tools are accessed through a standardized JSON-based interface that specifies input datasets, parameter settings, and expected outputs. This abstraction enforces reproducibility and allows analytical steps to be executed deterministically. Visualization operators generate publication-ready figures in both raster (PNG) and vector (PDF) formats, accompanied by tabular outputs (TSV). For exploratory inspection of high-dimensional datasets, LungChat provides an interactive heatmap interface using Morpheus and dedicated applications for single-cell label projection (Azimuth), visualization (ShinyCell) and exploration of spatial omics (Vitessce).

3.3.3 External Knowledge and Tools

LungChat integrates a set of API-backed external tools for functional enrichment, literature reasoning, clinical trials, and drug connectivity analysis. Functional enrichment is performed using ToppGene, enabling pathway, disease, and drug-target annotation of gene sets. Literature reasoning is supported through the PubTator3, EuropePMC APIs, allowing entity-aware retrieval across more than 11 million full-text PubMed Central articles and over 47 million PubMed abstracts and BioArxiv and MedArxiv preprints. Clinical trial information is retrieved via the ClinicalTrials.gov API. For all external queries, input terms undergo ontology-based expansion and synonym resolution prior to submission, improving recall while preserving biological specificity. Drug Connectivity analysis is performed through iLINCS, enabling connectivity mapping between disease-associated transcriptional signatures and chemical or genetic perturbation profiles.

3.3.4 Workflow Control, Context Management, and Safeguards

LungChat exposes a predefined set of workflow modes namely *Refine Question*, *Explore Data*, *Review Literature*, *Informed Analysis*, and *Test Hypothesis*. These constrain agent behavior, tool selection, and response synthesis to align with different stages of scientific inquiry. Metadata retrieval is performed in a targeted manner, loading only task-relevant fields into context to prevent overload during multi-step analytical workflows.

Short-term conversational state and intermediate workflow outputs are managed using a Redis-backed state store, enabling robust persistence across long analytical sessions. To ensure analytical integrity, the supervisor agent applies layered safeguards against prompt injection and adversarial inputs, including prompt-level validation and model-based content filtering for both text and voice inputs. These controls ensure that analytical workflows remain grounded in validated tools and clearly defined scientific objectives.

3.4 Multi-Agent Architecture

3.4.1 Agent Roles and Responsibilities

Table 1: Agent specifications.

Agent	Role	Analytical Scope
Supervisor	Query routing, result synthesis, code execution	Cross-modal coordination
Single-Cell	Differential expression, cell characterization	scRNA-seq transcriptomics
Spatial	Tissue organization, cell communication	Spatial transcriptomics
Research	Literature integration, enrichment, connectivity	External knowledge
Atlas	Ontology queries, vocabulary lookup	Domain databases

LungChat employs five specialized agents, each designed for a distinct aspect of multi-omics analysis (Table 1). The **Supervisor Agent** serves as the central orchestrator, determining optimal delegation strategies and possessing unique capabilities including dynamic analytical code execution

and direct access to aggregated expression data. The **Single-Cell Agent** specializes in cell-level transcriptomic analysis, operating on pre-processed datasets with computed statistics for rapid interactive exploration. The **Spatial Agent** handles tissue-level analyses including spatial gene expression mapping and cell-cell communication inference across multiple platforms. The **Research Agent** bridges internal data with external knowledge through literature search, functional enrichment analysis, and drug-gene connectivity mapping. The **Atlas Agent** provides structured access to domain-specific resources and controlled vocabularies.

3.4.2 Inter-Agent Communication and Execution Model

Agent coordination follows a stateless delegation model designed for reproducibility and modularity. Sub-agents execute independently without access to conversation history or other agents' intermediate reasoning; they receive only the delegated task specification and return self-contained responses. This design ensures that any sub-agent invocation produces identical outputs given identical inputs.

The supervisor implements parallel task execution for independent analytical operations. When a query decomposes into non-dependent sub-tasks, these execute concurrently, significantly reducing response latency. Dependent tasks execute sequentially with results from prior steps informing subsequent operations. Task execution follows a managed lifecycle with explicit states: *pending*, *in_progress*, and *completed*. Failed tasks trigger supervisor-mediated retry with parameter adjustment before graceful degradation with informative error reporting.

To support inter-agent collaboration and long-horizon workflows, LungChat employs a shared virtual filesystem serving as a common workspace for intermediate results and analysis outputs. When tool executions produce outputs exceeding a configurable size threshold, results are automatically materialized as filesystem outputs and replaced in context with lightweight file references. All agents have read and write access to the shared workspace, enabling downstream agents to inspect, extend, or refine outputs produced by upstream analyses.

The stateless delegation model serves a dual purpose beyond modularity: it implements context engineering principles that mitigate known limitations of long-context language model reasoning Liu et al. (2024). Each specialist agent operates in a scoped context window containing only its domain-relevant tools and instructions, preventing the cross-domain parameter leakage and instruction dilution that arise when a single agent must reason over all 51 tools simultaneously. Sub-agents are invoked without conversation history, ensuring that reasoning quality at turn n of a session is structurally independent of prior turns, a property we term *stale-context resistance*. This design aligns with recent findings that multi-agent collaboration through scoped worker contexts outperforms monolithic single-agent approaches on long-context tasks Zhang et al. (2024) and recommended practices for managing context degradation in agentic systems Anthropic (2025).

3.5 Tool-Augmented Reasoning

3.5.1 Tool Abstraction Layer

LungChat augments language model reasoning with a structured library of analytical tools organized into four functional operator classes: **Analysis Operators** perform statistical and computational procedures including ToppGene for functional enrichment testing and iLINCS for drug-gene perturbation connectivity analysis, alongside differential expression and pathway overlap operators; **Visualization Operators** generate publication-ready figures in both raster and vector formats; **Knowledge Operators** interface with external resources including biomedical literature databases, clinical trial registries, and gene annotation services; **Data Access Operators** provide structured access to consortium data APIs and curated omics datasets.

3.5.2 Hybrid Tool Selection Mechanism

With 51 tools across 4 operator classes, exhaustive evaluation of all tools for each query is computationally prohibitive and degrades response quality through irrelevant context. We implement a two-stage hybrid selection mechanism balancing broad recall with precise final selection.

Given a query q and tool library $\mathcal{T}$, the selection process is:

$$\mathcal{T}_{\text{selected}} = \text{LLM-Refine}\left(\text{Top-}K\left(\text{sim}_{\text{vec}}(q, \mathcal{T}) + \lambda \cdot \text{BM25}(q, \mathcal{T})\right), k_{\max}\right) \tag{1}$$

where sim_{vec} computes semantic embedding similarity and BM25 provides lexical keyword matching. In practice, $\lambda = 0.3$ balances semantic and lexical signals, and $K = 15$ defines the initial candidate pool. The candidate pool is then refined by a lightweight language model to select the final $k_{\max}$ tools (typically 3) based on query-tool relevance reasoning. This hybrid approach captures both semantic similarity for paraphrased queries and lexical overlap for specific technical terms, substantially reducing token consumption while maintaining selection accuracy.

3.6 Multi-Omics Data Integration

3.6.1 Data Modalities

LungChat integrates three complementary omics modalities through a unified query interface, as summarized in Table 2. **Single-cell RNA-seq** data comprises ¿37 datasets totaling approximately 2.8 million cells spanning fetal development, healthy adult tissue, and 15 disease contexts including idiopathic pulmonary fibrosis, COPD, and COVID-19, with pre-computed differential expression statistics and harmonized cell type annotations across common expression schemas. **Spatial transcriptomics** data includes IPF dataset across both Visium-HD and Xenium platform types capturing spatially-resolved gene expression with pre-computed cell-cell communication objects. **Pseudobulk expression** data provides aggregated views across cell types and conditions in a queryable warehouse with comprehensive differential expression signatures supporting cross-dataset comparisons.

3.6.2 Ontology-Grounded Metadata Harmonization

Metadata harmonization is performed by mapping heterogeneous author-provided annotations to standardized biomedical ontologies including CL, UBERON, MONDO, EFO, and HANCESTRO. This approach achieves robust resolution of heterogeneous annotations to canonical ontology terms, enabling consistent cross-dataset queries. All ontology mappings are versioned and documented to ensure consistent interpretation.

3.7 Direction-Aware Repurposing and Targeting (DART)

To enable cell-type-resolved therapeutic prioritization, we introduce DART, a method that distinguishes perturbations reversing disease-associated transcriptional programs from those that reinforce them. Unlike conventional connectivity analyses producing compound-level scores aggregated across heterogeneous cell populations, DART preserves cell-type resolution throughout analysis, allowing antagonistic effects across lineages to be explicitly identified.

Directional Scoring. Given a disease differential expression signature $\Delta_d(c) = \{(g_i, \log_2 \text{FC}_i^d)\}$ for cell type c and a perturbation-induced transcriptional signature $\Delta_p = \{(g_j, \log_2 \text{FC}_j^p)\}$, DART converts fold changes to binary directional indicators ($+1$ for upregulation, -1 for downregulation). For overlapping genes $G_{\text{overlap}} = G_d \cap G_p$, genes are classified by directional concordance:

$$G_{\text{reversed}} = \{g \in G_{\text{overlap}} \mid \text{sign}(\log_2 \text{FC}_g^d) \times \text{sign}(\log_2 \text{FC}_g^p) < 0\} \tag{2}$$

$$G_{\text{amplified}} = \{g \in G_{\text{overlap}} \mid \text{sign}(\log_2 \text{FC}_g^d) \times \text{sign}(\log_2 \text{FC}_g^p) > 0\} \tag{3}$$

DART computes two heuristic metrics: Pearson correlation on binary directions $\rho(c) = \text{corr}_{\text{Pearson}}(\text{dir}_d, \text{dir}_p)$ and reversal rate $R(c) = |G_{\text{reversed}}|/|G_{\text{overlap}}|$, where negative ρ and high R indicate therapeutic reversal. Perturbations are ranked by reversal rate (reversal mode) or its complement (amplification mode), enabling prioritization of compounds that oppose or reinforce disease programs. Directional stratification by disease-upregulated versus disease-downregulated genes enables cell-type-specific safety profiling. DART is implemented as a composable analysis operator operating on precomputed signatures and can be applied wherever matched disease and perturbation transcriptional profiles are available.

Algorithm 1 Query Decomposition and Execution

Require: Query q, Agent set $\mathcal{A}$, Tool library $\mathcal{T}$
Ensure: Response $\mathcal{R} = (\mathcal{V}, \mathcal{I}, \mathcal{A})$
plan $\leftarrow$ DECOMPOSE(q)
for task $\in$ plan.parallel_tasks **do**
    agent $\leftarrow$ SELECTAGENT(task)
    SPAWN(task, agent)
end for
for task $\in$ plan.sequential_tasks **do**
    result $\leftarrow$ EXECUTE(task)
    plan $\leftarrow$ UPDATEPLAN(plan, result)
end for
results $\leftarrow$ AWAIT(plan.all_tasks)
$\mathcal{R} \leftarrow$ SYNTHESIZE(results)
return $\mathcal{R}$

3.8 ITERATIVE SCIENTIFIC WORKFLOWS

3.8.1 PLANNING AND TASK DECOMPOSITION

Complex scientific queries require structured decomposition into executable sub-tasks. LungChat implements explicit planning through a task decomposition mechanism that transforms natural language queries into coordinated analytical workflows, see Algorithm 1.

The DECOMPOSE function employs few-shot prompting with examples of query-to-task mappings. SELECTAGENT uses rule-based routing based on task keywords with LLM fallback for ambiguous cases. The planning mechanism identifies task dependencies to maximize parallel execution. The supervisor can dynamically revise plans based on intermediate results - if initial analysis reveals unexpected patterns, subsequent tasks can be adjusted or additional analyses spawned without requiring user intervention.

3.8.2 PROVENANCE-TRACKED REPRODUCIBILITY AND MEMORY

Scientific analysis requires reproducibility beyond context. LungChat structures all outputs as persistent outputs with complete provenance: **Configuration Outputs** capture all input parameters, data filters, and analysis settings as JSON; **Visualization Outputs** include figures in multiple formats (PNG, PDF) with accompanying data tables (TSV); **Execution Traces** document the sequence of analytical operations, agent delegations, and intermediate results; **Session Reports** provide automated summaries of analytical workflows and key findings, see Appendix C for an example. Code, configurations and input files can downloaded and re-run by the user locally to reproduce results or the JSON directly supplied back to LungChat to replicate or modify prior analyses.

Beyond single-session analysis, LungChat maintains user-specific long-term analytical context across sessions, storing structured research state including previously analyzed genes, selected cell types, and references to generated outputs. Memory retrieval is scoped to analytical context and does not influence model parameters or tool execution logic, ensuring reproducibility while enabling cumulative scientific workflows. Extended analyses exceeding context capacity employ dynamic compression using a dedicated summarization module, preserving recent interactions while compressing prior context. Combined with filesystem-based eviction of large tool outputs, this strategy decouples analytical data volume from context length.

3.9 IMPLEMENTATION AND INFERENCE REGIME

LungChat operates entirely at inference time using frozen foundation models. No task-specific training, fine-tuning, or gradient updates are performed on any model components. All capabilities including multi-agent coordination, tool selection, and response synthesis emerge from prompt engineering, structured tool augmentation, and orchestration logic. The system employs Claude Sonnet 4.6 and GPT-4 class models with long-context windows ($\geq$128,000 tokens). Tool selection uses low-temperature decoding (set 0.1), while response synthesis uses moderate temperature (set

0.7). This design enables rapid capability updates through prompt and tool modifications without retraining, while leveraging the broad knowledge encoded in pre-trained foundation models.

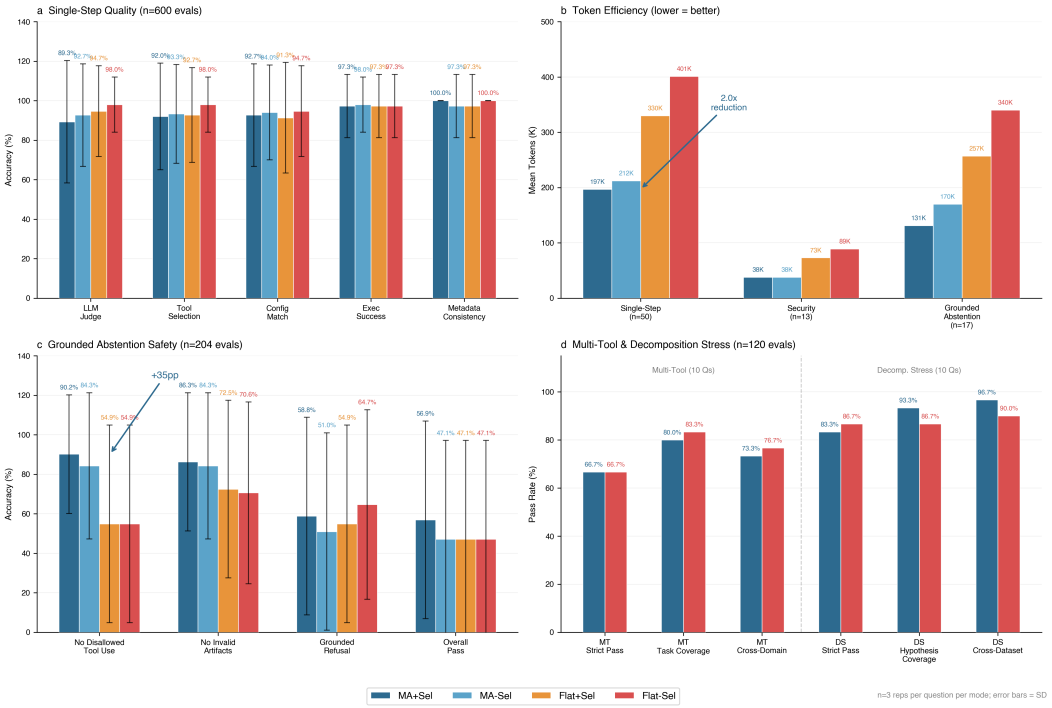

Figure 2: **Architecture ablation evaluation across all five tracks (1,080 evaluations).** (a) Single-step quality across four ablation configurations (50 queries $\times$ 3 reps $\times$ 4 modes $=$ 600 evaluations), (b) mean token consumption per query across three evaluation tracks (lower is better), (c) grounded abstention safety metrics (17 queries $\times$ 3 reps $\times$ 4 modes $=$ 204 evaluations), showing the +35pp improvement in disallowed-tool blocking for the multi-agent configuration, and (d) multi-tool orchestration and decomposition-stress results (20 queries $\times$ 3 reps $\times$ 2 modes $=$ 120 evaluations), demonstrating architecture-neutral quality on complex tasks.

4 Evaluation

To evaluate LungChat's ability to accurately plan, synthesize and execute multi-step workflows, we focused on pulmonary fibrosis, in which only a small number of candidate therapies exist. Existing therapies have focused largely on repression of activation, migration, proliferation of resident lung fibroblasts, which promote scarring and reduced lung function.

4.1 System Evaluation and Validation

4.1.1 Benchmark Design and Evaluation Framework

We evaluated LungChat using a comprehensive benchmark comprising 100 test cases across five evaluation tracks: 50 single-step analytical queries (analytical workflows, data visualization, metadata and knowledge queries), 13 security and adversarial queries (prompt injection, security exploitation, out-of-domain), 17 grounded abstention queries testing appropriate scope limitation, 10 multi-tool orchestration queries requiring cross-domain coordination, and 10 decomposition-stress queries testing scientific reasoning depth. To assess the contribution of the multi-agent architecture, we conducted ablation experiments across four configurations. Following current evaluation best practices Anthropic (2026), we employed both deterministic code-based evaluators (configuration correctness, execution success, metadata consistency, tool selection accuracy) and LLM-as-judge

scoring using GPT-4o (temperature 0) with a structured binary rubric covering biological accuracy, visual correctness, and refusal appropriateness. Each test case was executed across three independent trials (n=3 repetitions per case) to assess reliability.

4.1.2 System Performance and Architecture Ablation

To isolate the contribution of hierarchical multi-agent orchestration, we evaluated LungChat under four ablation configurations: (1) **MA+Sel** (production), the hierarchical architecture with hybrid tool selection; (2) **MA−Sel**, production hierarchy without tool selector; (3) **Flat+Sel**, a single agent with all 51 tools and merged domain prompts, retaining tool selector; and (4) **Flat−Sel**, a single agent with all tools, merged prompts, and no selector. Flat-agent baselines preserve all domain knowledge from specialist prompts, ensuring differences reflect architectural design rather than information asymmetry.

Figure 2 summarizes the ablation results across all five tracks (100 questions, 1,080 total evaluations). On single-step queries (Figure 2a), all four configurations achieve $\geq$89% LLM Judge accuracy, with Flat−Sel scoring highest (98.0%), confirming that single-step quality is architecture-neutral. The hierarchical architecture achieves a consistent $\sim$2$\times$ token reduction across all tracks (Figure 2b), reflecting the context-partitioning benefit of scoped tool pools described in Section 3.4.2. Grounded abstention (Figure 2c) reveals the largest architectural difference: MA+Sel blocks 90.2% of disallowed tool invocations versus 54.9% for flat architectures (+35pp). On complex multi-tool and decomposition-stress queries (Figure 2d), both architectures achieve comparable quality, confirming that analytical depth stems from domain engineering rather than hierarchy. Full per-track results are provided in Appendix F.

Adversarial testing validated the system's security posture: the system achieved 100% resistance to prompt injection attacks, developer mode requests, and instruction override attempts, with 100% blocking of system reconnaissance and environment variable extraction attempts. All analytical outputs are saved with complete provenance for deterministic re-execution (Appendix C).

4.2 Biological Validation on Established Disease Biology

4.2.1 Therapeutic Targets Identification

As a first evaluation, we explicitly asked LungChat to identify therapeutic targets for pulmonary fibrosis based on differential expression, pathway analysis, spatial location, and cell communication analyses. Our supervisor agent decomposed this query into six parallel sub-tasks delegated to all major specialized agents (Figure 3a,b).

These analyses successfully identify well-defined pulmonary fibrosis marker genes (e.g., IL13RA2, COL1A1) and enriched pathways of IPF (e.g., extracellular matrix deposition, P53 transcriptional regulation, VEGFA and TGF-β signaling; Figure 3c-e). Integrated spatial-based cell-communication analyses (Squidpy, CellChat) further nominated EREG and VEGFA signaling from aberrant basaloid cells (pulmonary fibrosis KRT5-/KRT17+ alveolar type 2 cells) to activated fibroblasts, as uniquely spatial enriched interactions in pulmonary fibrosis versus controls (Figure 3f). This result is supported by VEGFA/VEGFR2 signaling from unbiased global gene set enrichment (Figure 3d). LungChat performed drug repurposing analysis via iLINCS connectivity mapping combined with DART cell-type safety profiling. In addition to nominating FDA-approved IPF drugs (nintedanib and its primary target, PDGFRB), LungChat identified Saracatinib, a selective SRC kinase inhibitor, as the top therapeutic candidate (Figure 3g). DART cell-type impact analysis across 475 lung cellular signatures demonstrated that Saracatinib exhibits 90.9% reversal of pathogenic myofibroblast signatures (Pearson $r = -0.82$) and 87.5% reversal of alveolar fibroblast activation ($r = -0.75$), with an average fibroblast specificity of 54.8%. Off-target safety profiling identified potential risks in fetal/adult smooth muscle and Pericytes impacting dozen genes in each. Saracatinib demonstrated superior anti-fibrotic efficacy compared to nintedanib and pirfenidone in preclinical models Ahangari et al. (2022) and is currently being evaluated in a Phase 1b/2a clinical trial (STOP-IPF, NCT04598919). This therapeutic target identification was performed as a single natural-language prompt submitted to a fresh LungChat session with no prior conversational context or iterative prompt refinement; the complete execution trace, including all tool calls, parameters, and intermediate outputs, is provided in Appendix E.

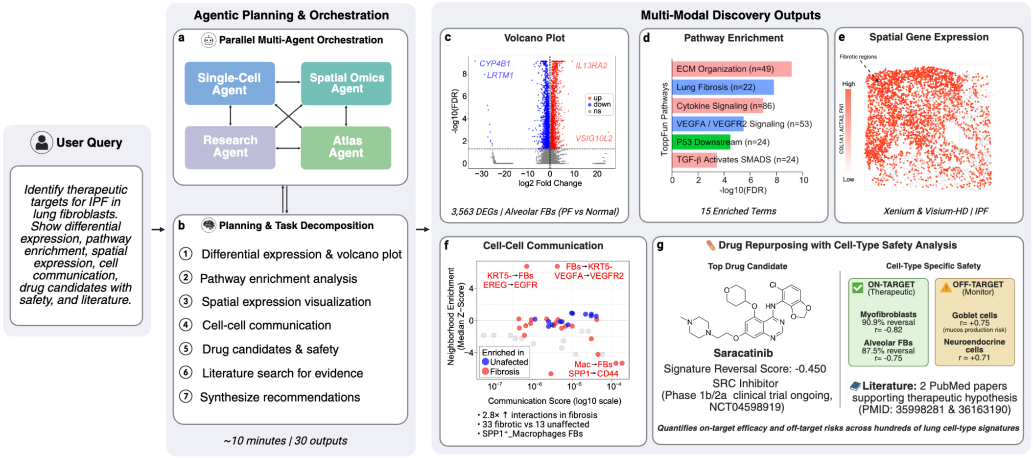

Figure 3: **Automated therapeutic target and drug prioritization for pulmonary fibrosis.** (a) LungChat parallel multi-agent orchestration with supervisor-coordinated task delegation. (b) Planning and task decomposition for reproducible workflows. (c) Ranking of up- and down-regulated genes in pulmonary fibrosis alveolar fibroblasts. (d) Gene-set enrichment (ToppGene) identifying known druggable pathways in pulmonary fibrosis. (e) Spatial transcriptomic localization of fibrotic markers. (f) Prediction of receptor-ligand interactions ranked by cell-type spatial proximity in pulmonary fibrosis versus controls (Xenium) cell-cell communication strength. (g) Drug repurposing with DART cell-type safety analysis: Saracatinib identified as top candidate. Off-target monitoring identified goblet cells and neuroendocrine cells.

As a separate evaluation, we asked LungChat to broadly predict drugs that would reverse disease gene expression profiles in any (separate) or all (union) cell populations from patients with COPD. This analysis identified Fluticasone Propionate, known commercially as Flonase, as among the top predicted anti-COPD drug reversal signatures in dendritic cells. Fluticasone Propionate is a synthetic trifluorinated corticosteroid which reduces airway inflammation in moderate-to-severe COPD in concert with β2-agonists, such as Fenoterol. Fenoterol was identified among the top 3 anti-COPD drug reversal signatures, when all cell type COPD DEGs were aggregated (union) by LungChat. DART predicted off-target reversal of fetal and adult normal gene expression profiles in fetal smooth muscle, neonatal secondary crest myofibroblasts and adult ionocytes. Given the rarity of ionocytes and relative recent nature of their functional characterization, such cellular impacts would likely go unnoticed in conventional off-target screening cellular platforms.

4.2.2 Evaluating Regulatory Mechanisms

In addition to the discovery and safety evaluation of promising therapeutic candidates, LungChat can deeply evaluate regulatory mechanisms from a diverse repertoire of omics-focused analyses and the scientific literature. To assess this capability, LungChat was tasked to evaluate the potential mechanisms of action of a new drug target under current FDA evaluation. Lysophosphatidic acid receptor (LPAR) inhibitors target TGF-beta signaling pathway members, which are well-documented to be up-regulated in pulmonary fibrosis. To determine the mechanism of action of these inhibitors, LungChat initiated a multi-step analysis to: 1) identify candidate drugs in iLINCS, 2) determine which cell types express LPAR1, 3) identify its spatial niches, 4) determine which cell-type disease programs could be reversed by drug treatment and to identify which transcriptional regulators were upstream of pharmacologically reversed genes (transcriptionally). These analyses confirmed expression of LPAR1 in all adult lung fibroblast and pericyte populations, confirmed spatial expression of LPAR1 with activated fibroblasts in the IPF niche, identified expected and novel IPF cellular impacts (reversal of IPF expression in fibroblasts and myeloid cells, respectively) and identified SMAD3 as the upstream transcriptional regulator of LPAR inhibitor targets genes (upregulation of SERPINE1). However, DART identified a dozen of potential off-target effects, including in presumably healthy smooth muscle, lymphatic endothelial, submucosal gland, deutrosomal, fetal Schwann

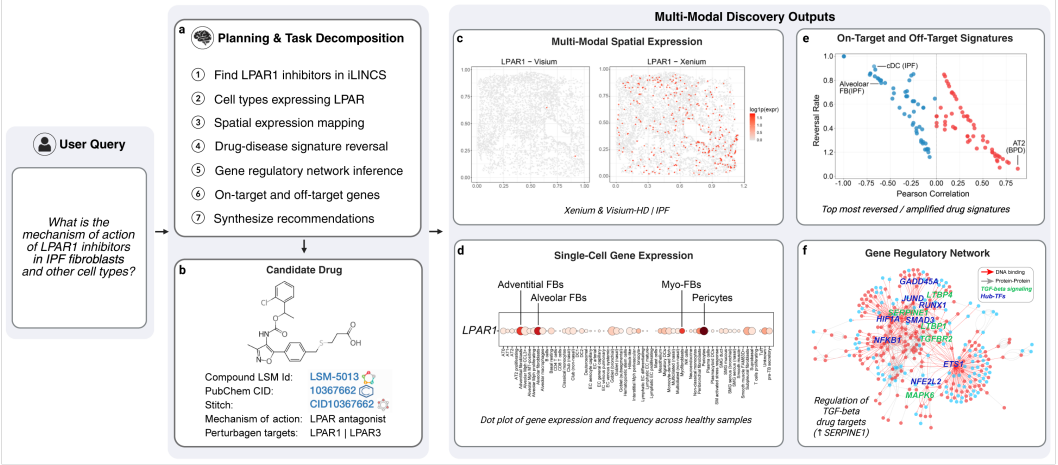

Figure 4: **Resolving the regulatory mechanisms and cell targets of new therapies.** (a) Planning and task decomposition to define diverse mechanisms of action for LPAR1 inhibitors in pulmonary fibrosis. (b) Identification of LPAR1 inhibitors in iLINCS with in vitro gene expression responses. (c) Comparison of the spatial distribution LPAR1 in Xenium and Visium-HD (same slide - Spatial Agent). (d) Expression and frequency of LPAR1 across adult healthy lung cell populations (HLCA - Single Cell Agent). (e) DART prediction of the top impacted lung disease, technology and developmental perturbations (Omics Data Repository), reversed or further amplified by in vitro LPAR agonist treatment (iLINCS). (f) Inferred gene regulatory network from annotated transcription factor (TF) target and pathway interactions (NetPerspective), with highlighted major TF nodes and annotated TGF-beta signaling responsive genes (ToppGene).

cell and myofibroblasts with LPAR1 inihibition. Thus, LungChat is able to autonomously define complex regulatory mechanisms underlying drug treatment, including known and novel predicted interactions.

Table 2: LungChat data and system coverage.

Resource	**Coverage**
Single-cell datasets	39 (2.8M cells)
Spatial datasets	1 (Visium-HD, Xenium)
Disease contexts	15
Specialized tools	51
Precomputed DEG signatures	579

5 Conclusion

We present LungChat, a conversational multi-agent AI framework for integrated multi-omics analysis that enables transparent, end-to-end scientific workflows across single-cell, spatial, and pseudobulk transcriptomic modalities without requiring manual scripting or domain-specific expertise. Architecture ablation across five evaluation tracks (1,080 evaluations) demonstrates that analytical quality is architecture-neutral, while the multi-agent hierarchy provides $\sim 2\times$ token efficiency and a +35 percentage-point improvement in grounded abstention, validating context-engineering principles for production deployment.

Through a series of therapeutically focused case studies, we demonstrate that LungChat can faithfully recover established disease pathobiology, regulatory disease processes, contrast disease and healthy spatial niches, and identify established antifibrotic therapies. Beyond recapitulating known biology, we introduce Direction-Aware Repurposing and Targeting (DART), a cell-type-resolved method to predict therapeutic reversal of disease signatures or drug treatments based on altered

transcriptional programs. Applied to IPF, DART identified Saracatinib, a selective SRC kinase inhibitor, as the top therapeutic candidate, demonstrating 90.9% reversal of pathogenic myofibroblast signatures (Pearson $r = -0.82$) which was validated by its advancement to clinical trials. Similar confirmatory findings were observed in COPD using DART, opening the door for prospective studies evaluating newly nominated compounds.

More broadly, DART's cell-type-resolved analysis can identify conserved disease or developmental gene programs, technology-specific biases (e.g., Xenium versus Visium-HD), or identify analogous cell types across studies that may have been mislabeled. These analyses have the potential to significantly optimize drug-discovery workflows by defining the precise molecular pathways, cell populations, and disease conditions positively or adversely impacted by an intentional perturbation. By revealing asymmetric perturbation effects across cellular populations, DART enables safety-aware hypothesis generation and highlights off-target risks that would be masked by bulk analyses.

Importantly, LungChat emphasizes analytical correctness and reproducibility over autonomous hypothesis generation, with all analyses producing provenance-tracked outputs supporting deterministic re-execution. While instantiated for lung biology, the architecture is tissue-agnostic, with modular agents and standardized data schemas enabling extension to other organs and multi-omics domains. LungChat provides a scalable, cost-effective framework for integrated data analysis with grounded, reproducible biological interpretation. In summary, our proposed framework demonstrates how conversational multi-agent systems can function as practical scientific interfaces that close the loop between reasoning and experimentation. By aligning foundation models with validated analytical tools and reproducible execution, LungChat provides a blueprint for safety-aware, transparent AI systems that augment, rather than replace biological discovery.

Use of Large Language Models: This manuscript was written by the authors with assistance from large language models for grammar improvement, sentence clarity, and writing refinement. All scientific content, results, analyses, and conclusions were developed by the authors, who take full responsibility for the accuracy and validity of all claims.

REFERENCES

Farida Ahangari, Christine Becker, Daniel G Foster, Maurizio Chioccioli, Meghan Nelson, Keriann Beke, Xing Wang, Aurelien Justet, Taylor Adams, Benjamin Readhead, et al. Saracatinib, a selective src kinase inhibitor, blocks fibrotic responses in preclinical models of pulmonary fibrosis. *American Journal of Respiratory and Critical Care Medicine*, 206(12):1463–1479, Dec 2022. doi: 10.1164/rccm.202010-3832OC.

Anthropic. Building effective agents. `https://www.anthropic.com/research/building-effective-agents`, 2025. Accessed: 2026-03-28.

Anthropic. Demystifying evals for ai agents. `https://www.anthropic.com/engineering/demystifying-evals-for-ai-agents`, Jan 2026. Accessed: 2026-03-28.

Katy Börner, Philip D Blood, Jonathan C Silverstein, Matthew Ruffalo, Rahul Satija, Sarah A Teichmann, Gloria J Pryhuber, Ravi S Misra, Jeffrey M Purkerson, Jean Fan, et al. Human biomolecular atlas program (hubmap): 3D human reference atlas construction and usage. *Nature Methods*, pp. 1–16, 2025.

Jing Chen, Eric E Bardes, Bruce J Aronow, and Anil G Jegga. ToppGene suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Research*, 37:W305–W311, 2009.

Nathan Gaddis, Joshua Fortriede, Minzhe Guo, Eric E Bardes, Michal Kouril, Scott Tabar, Kevin Burns, Maryanne E Ardini-Poleske, Stephanie Loos, Daniel Schnell, et al. LungMAP portal ecosystem: systems-level exploration of the lung. *American Journal of Respiratory Cell and Molecular Biology*, 70(2):129–139, 2024.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Peng He, Kyungtae Lim, Dawei Sun, Jan Patrick Pett, Quitz Jeng, Krzysztof Polanski, Ziqi Dong, Liam Bolt, Laura Richardson, Lira Mamanova, et al. A human fetal lung cell atlas uncovers proximal-distal gradients of differentiation and key regulators of epithelial fates. *Cell*, 185(25): 4841–4860, 2022.

Kexin Huang, Payal Chandak, Qianwen Wang, Shreyas Havaldar, Akhil Vaid, Jure Leskovec, Girish N Nadkarni, Benjamin S Glicksberg, Nils Gehlenborg, and Marinka Zitnik. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30(12):3601–3613, 2024.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Junze Zhang, Yin Di, et al. Biomni: A general-purpose biomedical ai agent. biorxiv, 2025.

Yoshitaka Inoue, Tianci Song, and Tianfan Fu. Drugagent: Explainable drug repurposing agent with large language model-based reasoning. *arXiv preprint arXiv:2408.13378*, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Suoqin Jin, Christian F Guerrero-Juarez, Lihua Zhang, Ivan Chang, Raul Ramos, Chen-Hsiang Kuan, Peggy Myung, Maksim V Plikus, and Qing Nie. Inference and analysis of cell-cell communication using CellChat. *Nature Communications*, 12(1):1088, 2021.

Oussama Kharouiche, Aris Markogiannakis, Xiao Fei, Michail Chatzianastasis, and Michalis Vazirgiannis. Cell2Text: Multimodal LLM for generating single-cell descriptions from RNA-seq data. *arXiv preprint arXiv:2509.24840*, 2025.

Guangyuan Li, Baobao Song, Harinder Singh, VB Surya Prasath, H Leighton Grimes, and Nathan Salomonis. Decision level integration of unimodal and multimodal single cell data with scTriangulate. *Nature Communications*, 14(1):406, 2023.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024a.

Yen-Chun Lu, Ashley Varghese, Rahul Nahar, Hao Chen, Kunming Shao, Xiaoping Bao, and Can Li. scChat: A large language model-powered co-pilot for contextualized single-cell rna sequencing analysis. *BioRxiv*, pp. 2024–10, 2024b.

Nikita Mehandru, Amanda K Hall, Olesya Melnichenko, Yulia Dubinina, Daniel Tsirulnikov, David Bamman, Ahmed Alaa, Scott Saponas, and Venkat S Malladi. BioAgents: Bridging the gap in bioinformatics analysis with multi-agent systems. *Scientific Reports*, 15(1):39036, 2025.

L Mitchener, A Yiu, B Chang, M Bourdenx, T Nadolski, A Sulovari, EC Landsness, DL Barabasi, S Narayanan, N Evans, et al. Kosmos: An AI scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*, 2025.

Giovanni Palla, Hannah Spitzer, Michal Klein, David Fischer, Anna Christina Schaar, Louis Benedikt Kuemmerle, Sergei Rybakov, Ignacio L Ibarra, Olle Holmberg, Isaac Virshup, et al. Squidpy: a scalable framework for spatial omics analysis. *Nature Methods*, 19(2):171–178, 2022.

Marcin Pilarczyk, Mehdi Fazel-Najafabadi, Michal Kouril, Behrouz Shamsaei, Juozas Vasiliauskas, Wen Niu, Naim Mahi, Lixia Zhang, Nicholas A Clark, Yan Ren, et al. Connecting omics signatures and revealing biological mechanisms with iLINCS. *Nature Communications*, 13(1):4678, 2022.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, et al. CZ CELLxGENE discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 2025.

Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The human cell atlas. *eLife*, 6:e27041, 2017.

Summer Rosonovski, Maria Levchenko, Rajat Bhatnagar, Umamageswari Chandrasekaran, Lynne Faulk, Islam Hassan, Matt Jeffryes, Syed Irtaza Mubashar, Maaly Nassar, Madhumiethaa Jayaprabha Palanisamy, et al. Europe pmc in 2023. *Nucleic Acids Research*, 52(D1):D1668–D1676, 2024.

Orit Rozenblatt-Rosen, Michael JT Stubbington, Aviv Regev, and Sarah A Teichmann. The human cell atlas: from vision to reality. *Nature*, 550(7677):451–453, 2017.

Anna C Schaar, Alejandro Tejada-Lapuerta, Giovanni Palla, Robert Gutgesell, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Mojtaba Bahrami, et al. Nicheformer: A foundation model for single-cell and spatial omics. *BioRxiv*, pp. 2024–04, 2024.

Moritz Schaefer, Peter Peneder, Daniel Malzl, Salvo Danilo Lombardo, Mihaela Peycheva, Jake Burton, Anna Hakobyan, Varun Sharma, Thomas Krausgruber, Celine Sin, et al. Multimodal learning enables chat-based exploration of single-cell data. *Nature Biotechnology*, pp. 1–11, 2025.

Shawyon P Shirazi, Nicholas M Negretti, Christopher S Jetter, Alexandria L Sharkey, Shriya Garg, Meghan E Kapp, Devan Wilkins, Gabrielle Fortier, Saahithi Mallapragada, Nicholas E Banovich, et al. Bronchopulmonary dysplasia with pulmonary hypertension associates with semaphorin signaling loss and functionally decreased FOXF1 expression. *Nature Communications*, 16(1): 5004, 2025.

Lisa Sikkema, Ciro Ramírez-Suástegui, Daniel C Strobl, Tessa E Gillett, Luke Zappia, Elo Madissoon, Nikolay S Markov, Laure-Emmanuelle Zaragosi, Yuge Ji, Meshal Ansari, et al. An integrated cell atlas of the lung in health and disease. *Nature Medicine*, 29(6):1563–1577, 2023.

Matthew L Speir, Aparna Bhaduri, Nikolay S Markov, Pablo Moreno, Tomasz J Nowakowski, Irene Papatheodorou, Alex A Pollen, Brian J Raney, Lucas Seninge, W James Kent, et al. UCSC cell browser: visualize your single-cell data. *Bioinformatics*, 37(23):4578–4580, 2021.

Houcheng Su, Weicai Long, and Yanlin Zhang. BioMaster: Multi-agent system for automated bioinformatics analysis workflow. *BioRxiv*, pp. 2025–01, 2025.

Meenakshi Venkatasubramanian, Kashish Chetal, Daniel J Schnell, Gowtham Atluri, and Nathan Salomonis. Resolving single-cell heterogeneity from hundreds of thousands of cells through sequential hybrid clustering and NMF. *Bioinformatics*, 36(12):3773–3780, 2020.

Chih-Hsuan Wei, Alexis Allot, Po-Ting Lai, Robert Leaman, Shubo Tian, Ling Luo, Qiao Jin, Zhizheng Wang, Qingyu Chen, and Zhiyong Lu. Pubtator 3.0: an ai-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Research*, 52(W1):W540–W546, 2024.

Haotian Zhang, Yu H Sun, Wenxing Hu, Xu Cui, Zhengyu Ouyang, Derrick Cheng, Xinmin Zhang, and Baohong Zhang. CompBioAgent: An llm-powered agent for single-cell rna-seq data exploration. *BioRxiv*, pp. 2025–03, 2025.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tatsunori Hashimoto, and Daniel Fried. Chain of agents: Large language models collaborating on long-context tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

APPENDIX

A DATASET CURATION AND PREPROCESSING

A.1 SINGLE-CELL RNA-SEQ

Four independent single-cell RNA-Seq collections were integrated into LungChat's corpus of initial available datasets. The largest of these dataset collections consist of 2.2 million cells from 486 lungs from Human Lung Cell Atlas (HLCA) version 2 Sikkema et al. (2023) object in CellxGene (accession: 6f6d381a-7701-4781-935c-db10d30de293 https://cellxgene.cziscience.com/collections/6f6d381a-7701-4781-935c-db10d30de293). HLCA is a collection of reprocessed raw droplet sequencing data 35 single-cell studies, spanning 14 distinct disease entities. To efficiently assess and analyze this data, the R library SuperCell v.1.0 was used to produce 50,520 metadata cells from the HLCA, separately for each individual donor and HLCA consortia annotated cluster (ann_finest_level). A lung cell atlas of fetal development was also obtained from CellxGene (accession: 2d2e2acd-dade-489f-a2da-6c11aa654028 https://cellxgene.cziscience.com/collections/2d2e2acd-dade-489f-a2da-6c11aa654028), spanning 5-22 post-conception weeks He et al. (2022). Two additional LungMAP consortia datasets were provided to LungChat, comprising infant lungs at the time of death from bronchopulmonary dysplasia or other causes (controls) (LungMAP.net accession: LMEX0000004400 https://www.lungmap.net/dataset/?dataset_id=LMEX0000004400, LMEX0000004401 https://www.lungmap.net/dataset/?dataset_id=LMEX0000004401) Shirazi et al. (2025). Cell annotations for each dataset include author curated at different levels of resolution and Cell Ontology unique IDs. All datasets are formatted by LungChat pre-processing as anndata H5AD objects, from source H5AD or RDS, using developed automated processing workflows (GitHub). This workflow retains author provided UMAP coordinates (averaged for metacells) and harmonized sample metadata (CellxGene). The preprocessing workflow further produces normalized expression using the scanpy python framework (natural log of counts per 10,000 reads), translates Ensembl primary identifiers to gene symbols (Ensembl 100), computes cell population marker gene and covariate differential gene expression versus appropriate control conditions (wilcoxon, FDR corrected) and retains DEGs with fold > 1.2 and Mann-Whitney U test $p < 0.05$ (FDR adjusted) in uns["rank_genes_groups"]. Study-specific gene signatures are stored for later comparison in LungChat. Each anndata object is augmented with a database of protein-protein and protein-DNA interactions from the AltAnalyze NetPerspective database Venkatasubramanian et al. (2020), for fast retrieval and display of inferred gene regulatory networks in LungChat.

A.2 SPATIAL TRANSCRIPTOMICS

Image-based (10x Genomics Xenium) and whole transcriptome (10x Genomics Visium-HD) spatial transcriptomics from a collection of human healthy and IPF lungs were obtained from the Gene Expression Omnibus database (GSE276945 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE276945). Each patient sample corresponds to a region of interest (RO1), tiled on the Xenium (343 gene probes) or Visium slide (4,571 expressed genes, GSM8505452 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM8505452). For Xenium samples (n=35), author generated cell segmentation (1.6 million cells), curated cell annotation (Final_CT) and niche annotation (TNiche and CNiche) were retained, along with sample level metadata (i.e., sex, age). Segmented cell annotations were augmented with Robust Cell Type Decomposition (RCTD) from the spacexr package that was applied to derive supervised cell annotations for RCTD-classified singlets using the human CellRef transcriptome as a reference. For Visium-HD, as no author annotation for 8μm spots were provided, clusters, the same RCTD protocol was applied for Visium spots, to generate annotations.

To derive hybrid multi-modal cell annotations from the integration of matching Xenium and Visium samples, we reoriented, scaled and aligned Xenium segmented map (query) to Visium (reference) coordinates, considering unique cell/spot correspondences (dplyr package in R). For downstream integrated analyses, all cell/spots were downsampled to those with 1:1 correspondences using the find nearest neighbor function (FNN) package and get.knnx() function. Mapping was verified using COL1A1 expression, which was highly expressed by both plat-

forms. Cluster annotations were mapped across 1:1 mapped cells/spots across platforms to define the most stable cell populations using a new spatial library within the scTriangulate package (https://github.com/frankligy/scTriangulate). scTriangulate Li et al. (2023) uses a game-theory approach to compare distinct stability metrics (maker gene expression, reclassification) for each single-cell cluster, separately for each modality evaluated. In this instance, all Xenium and Visium cell or niche clusters were evaluated using the integrated object,with RNA values and spatial coordinates from each platform considered as separate modalities. A final set of 29 multimodal stable clusters were defined from this analysis and stored in the integrated RDS object for LungChat analysis.

To infer cell-cell communication networks, we applied the python package Squidpy to assess cell-type spatial proximity via neighborhood enrichment (z-score) and CellChat to infer cell-cell and receptor-ligand interactions. These analyses were pre-computed using the Xenium dataset considering control (less affected) and IPF (more affected) samples. Both Squidpy and CellChat were applied to different annotation layers, namely author (Final_CT), cellular niches (CNiche, TNiche) and Final_CT annotations that subdivide individual niches. Computed results were stored in Seurat RData objects for direct access by LungChat.

B Tool Reference and Output Specifications

B.1 Tool Scope

This appendix provides a capability-level reference of analytical, external API, and knowledge tools used by the Supervisor, Single-Cell, Spatial, and Research agents (LungMAP agent excluded).

B.2 Capability-Level Tool Catalog

Table B1: Capability-level reference of analytical, external API, and knowledge tools used by the Supervisor, Single-Cell, Spatial, and Research agents (LungMAP agent excluded), including output formats and descriptions.

Tool	Output format	Description
Supervisor Agent Tools		
`query_pseudobulk_expression_data_tool`	JSON	Query pseudobulk omics warehouse (DEG lookup, aggregation, comparisons, signature overlap, drug–cell-type impact)
`execute_code_tool`	JSON	Run Python or R in sandbox
Single-Cell (scRNA-seq) Agent Tools		
`plot_heatmap`	JSON, PDF, PNG, TSV	Clustered heatmap of gene expression across cell types/conditions.
`plot_radar`	JSON, PDF, PNG	Radar/spider chart of cell type composition across conditions.
`plot_cell_frequency`	JSON, PDF, PNG	Box/violin of cell frequencies per donor; Mann-Whitney U.
`plot_volcano`	JSON, PDF, PNG	Volcano: log2FC vs FDR; top genes labeled.
`plot_dotplot`	JSON, PDF, PNG	Dot plot: % expressing (size), mean expression (color) across cell types.

Tool	Output format	Description
plot_violin	JSON, PDF, PNG	Violin of one gene expression across cell types/conditions.
plot_venn	JSON, PDF, PNG	Venn diagram of gene set overlaps (2–3 sets).
plot_upset_genes	JSON, PDF, PNG	UpSet plot for 4+ gene set overlaps.
plot_umap	JSON, PDF, PNG	UMAP 2D embedding; optional gene overlay.
plot_network	JSON, PDF, PNG	Gene–gene interaction/regulatory network.
plot_gene_correlation	JSON, PDF, PNG, TSV	Pearson correlation of genes with query gene.
generate_stats	JSON, TSV	Export DEG/marker statistics.
Spatial Agent Tools		
plot_spatial	JSON, PDF, PNG	Spatial plot of cell annotations (Final_CT, TNiche, CNiche, etc.) on coordinates.
plot_visualize_celltypes_spatial_distribution	JSON, PDF, PNG	Multi-panel spatial distribution of cell types across samples/groups.
plot_compare_spatial_rna_with_histology_quadpanel	JSON, PDF, PNG	4-panel: Visium-HD, Xenium, H&E, DAPI.
plot_umap	JSON, PDF, PNG	UMAP colored by annotation (spatial data).
plot_visualize_gene_expression_spatial	JSON, PDF, PNG	Gene expression on spatial coordinates (Xenium/Visium).
plot_barchart_top_correlated_visium_xenium_rna	JSON, PDF, PNG	Bar chart of top correlated genes across Visium/Xenium.
plot_evaluate_cluster_stability_report	JSON, PDF, PNG	Report on annotation stability across sources.
rank_celltype_marker_genes_by_specificity	JSON, TSV	Rank cell types by marker specificity (t-test/FDR).
plot_celltype_marker_volcano	JSON, PDF, PNG	Volcano of marker genes for one cell type.
plot_compare_markers_across_platforms_dotplot	JSON, PDF, PNG	Dot plot comparing markers across platforms.
plot_evaluate_annotation_concordance_dotplot	JSON, PDF, PNG	Dot plot of annotation concordance.
plot_scatter_matched_xenium_visium_spots	JSON, PDF, PNG	Scatter of matched Xenium–Visium spot expression.
plot_differential_expression_volcano	JSON, PDF, PNG	Volcano for spatial DEGs between groups.
plot_cellchat_chord_diagram	JSON, PDF, PNG	CellChat chord diagram of signaling.

Tool	Output format	Description
plot_cellchat_heatmap	JSON, PDF, PNG	CellChat signaling strength heatmap.
plot_cellchat_signaling_pathway_network	JSON, PDF, PNG	CellChat pathway network.
extract_cellchat_interactions_subset	JSON, TSV	Extract subset of CellChat interactions.
plot_squidpy_neighbour_enrichment_heatmap	JSON, PDF, PNG	Squidpy neighborhood enrichment heatmap.
plot_squidpy_celltype_neighborhood_enrichment	JSON, PDF, PNG	Squidpy cell type neighborhood enrichment.
plot_identify_significant_proportion_changes	JSON, PDF, PNG, TSV	Proportion changes between groups.
plot_compare_different_cell_annotations_heatmap	JSON, PDF, PNG	Heatmap comparing two annotation columns.
plot_identify_celltypes_enriched_in_niches	JSON, PDF, PNG	Cell types enriched in niches.
plot_squidpy_compare_neighborhood_changes	JSON, PDF, PNG	Compare Squidpy neighborhoods across groups.
identify_unique_cellchat_interactions_spatial	JSON, TSV	Unique CellChat interactions per group.
plot_compare_communication_profiles	JSON, PDF, PNG	Compare CellChat communication between groups.
Research Agent Tools		
gene_functional_enrichment_tool	JSON, TSV	ToppGene enrichment (20 categories).
literature_search_tool	JSON	PubMed + Europe PMC; ranking, preprints, entity extraction. Returns papers, insights.
clinical_trials_search_tool	JSON	ClinicalTrials.gov v2; term expansion (MONDO/EFO). Search or retrieve by NCT ID.
drug_connectivity_tool	JSON, PDF, PNG, TSV	iLINCS: signature connectivity, compound investigation, gene knockout, similar compounds.
web_search_tool	JSON	Google Serper API; fallback for non-biomedical queries. Returns title, snippet, link.
think_tool	JSON	Record strategic reflection (progress, gaps, next steps). Returns confirmation.

C PROVENANCE-TRACKED CONFIGURATION EXAMPLE


```
{
  "adata_file": "HLCA_full_superadata_v3_norm_log_deg.h5ad",
  "plot_title": null,
  "cell_type": "Alveolar fibroblasts",
  "disease": "pulmonary fibrosis",
  "top_n": 8,
  "direction": "regulated",
  "cell_type_index": "ann_finest_level",
  "covariate_index": "disease",
  "plot_type": "volcano",
  "thread_id": "d8b70f18-0b56-45fc-a13d-4479e687d2cd",
  "base_url": "http://chat.lungmap.net/lungchat/public/output",
  "tool": "volcano",
  "timestamp": "2026-02-09T07:07:41.925473+00:00",
  "task": "scRNA-seq"
}
```


Figure C1: **Provenance-tracked configuration output for reproducible tool execution.** Example JSON payload automatically saved by LungChat for a differential-expression volcano plot request (dataset, cell type, disease, direction, and tool parameters), enabling deterministic re-execution and auditability of analysis settings as described in Section 4.1.3.

D CAPABILITY COMPARISON WITH RELATED SYSTEMS

Table D1: Capability-level comparison with related conversational and agentic biomedical AI systems.

Capability	LungChat	Frontier LLMs	CellWhisperer	CompBioAgent / scChat	BioAgents	Biomni
Conversational scientific interface	Strong	Strong	Strong	Strong	Moderate	Strong
Executable omics workflows	Strong	Limited	Limited	Moderate	Moderate	Strong
Multi-agent orchestration	Strong	Limited	Limited	Limited	Strong	Strong
Cross-dataset integration	Strong	Limited	Moderate	Moderate	Moderate	Strong
Single-cell + spatial integration	Strong	Limited	Limited	Limited	Limited	Moderate
Cell-type–resolved drug / perturbation analysis	Strong	Limited	Limited	Limited	Limited	Moderate
Evidence Ranked Literature Mining	Strong	Moderate	Limited	Limited	Moderate	Strong
Provenance-tracked outputs	Strong	Limited	Limited	Limited	Moderate	Moderate
Persistent analytical memory	Strong	Limited	Limited	Limited	Limited	Moderate
Domain-specific depth in lung multi-omics	Strong	Limited	Limited	Limited	Limited	Limited
General-purpose biomedical breadth	Moderate	Strong	Limited	Limited	Moderate	Strong

Direct quantitative benchmarking against competing systems is precluded by non-overlapping analytical capabilities: the end-to-end workflows evaluated here, spanning differential expression, functional enrichment, drug connectivity screening, cell-type–resolved perturbation analysis (DART), and spatial transcriptomics, require integrated tool access that no single competing system currently supports. While several systems offer public deployments (Biomni, CellWhisperer, CompBioAgent), each addresses a subset of these capabilities (Table D1). We therefore provide this qualitative capability comparison alongside a controlled single-agent vs. multi-agent architecture ablation (Section 4).

E DART CASE STUDIES

This appendix provides comprehensive execution provenance for the two DART case studies described in Section 4.2.1: (1) the IPF Saracatinib discovery, executed as a single-prompt, zero-shot workflow; and (2) the COPD Fluticasone Propionate discovery, executed as a single-prompt, cell-type-resolved zero-shot workflow with per-cell-type iLINCS screening. Both case studies were conducted in fresh sessions with no prior conversation history. All data, tables, and narrative text below are derived directly from the LungChat conversation exports and LangSmith agent execution traces. Figure placeholders reference generated outputs archived with the session.

E.1 IPF: SARACATINIB DISCOVERY

Supplementary E1. IPF DART Case Study: Single-Prompt Zero-Shot Therapeutic Discovery

User Prompt (verbatim).

> *Identify therapeutic targets and drug repurposing candidates for IPF based on cell-type–specific disease programs in fibroblast populations most associated with fibrosis, comparing diseased and healthy states within those cells, and how these manifest across lung cell types and tissue context, including safety. Prioritize drugs that reverse this program, and for the top candidate, assess its cell-type-specific effects across the lung.*

Supervisor Decomposition. LungChat's supervisor agent decomposed this single natural-language prompt into six coordinated sub-tasks, executed across three specialized agents with parallel dispatch for independent operations:

1. **Define the disease program:** DEGs in IPF fibroblasts vs. healthy controls (HLCA)
2. **Drug repurposing:** iLINCS screen for compounds reversing the fibroblast signature
3. **Pathway enrichment:** Characterize the fibrotic program biologically (ToppGene)
4. **Cell-type impact:** Map the top drug candidate across all lung cell types (DART)
5. **Spatial context:** Fibroblast niches and ligand-receptor signaling in IPF tissue
6. **Safety synthesis:** Integrate all findings into a prioritized report

Steps 1, 3, and 5 were dispatched in parallel. Steps 2 and 4 were executed sequentially after Step 1 completed.

Complete Execution Trace.

Step	Agent	Tool(s)	Operation
1	Supervisor	`decompose`	Parses user query into 6 sub-tasks; identifies parallelizable branches
2	Single-Cell	`generate_stats`	Differential expression: Alveolar Fibroblasts, IPF vs. control (HLCA). Produces 3,563 DEGs (FDR $<$ 0.05); 56 up (logFC $>$ 0.5), 55 down (logFC $<$ -0.5) at stringent thresholds
3	Research	`gene_functional_enrichment_tool`	ToppGene enrichment on 56 upregulated and 55 downregulated gene sets. Returns GO Biological Process, Pathway, Disease, and Drug annotations
4	Spatial	`plot_visualize_celltypes_spatial_distribution`, `plot_compare_communication_profiles`	Spatial distribution of fibroblast subtypes (Xenium); CellChat + Squidpy interaction analysis for Activated_Fibrotic_FBs (IPF vs. control). Returns 46 significant interactions, 33 unique to IPF
5	Research	`drug_connectivity_tool`	iLINCS connectivity analysis: 3,563-gene alveolar fibroblast IPF signature screened against 2,383 LINCSCP compounds in reversal mode. Returns ranked drug list
6	Supervisor	`drug_connectivity_tool` (investigate)	Retrieves Saracatinib's own iLINCS compound perturbation signature (976 genes) for downstream DART analysis
7	Supervisor	`query_pseudobulk_expression_data_tool` (`drug_cell_type_impact`)	Executes DART: correlates Saracatinib signature against 475 lung cell-type/disease signatures across HLCA, BPD_Sun, and HCA_fetal datasets
8	Supervisor	Synthesis	Aggregates all outputs; generates integrated therapeutic report with efficacy, spatial context, and safety profiling

Step 2 Output: Fibroblast Disease Program. Analysis of alveolar fibroblasts from the HLCA identified a robust fibrotic gene expression signature comprising **3,563 significant DEGs** (FDR < 0.05). The upregulated program is dominated by extracellular matrix remodeling, cell adhesion, and pro-inflammatory signaling; the downregulated program indicates loss of fibroblast detoxification and quiescence capacity.

Top Upregulated Genes (Active Fibrotic Program):

Gene	Log2FC	Functional Role
IL13RA2	$+9.69$	IL-13 decoy receptor; fibrosis amplifier
THY1	$+6.54$	Fibroblast activation marker (CD90)
NFATC2	$+5.63$	Transcription factor; myofibroblast activation
CPXM1	$+5.59$	Extracellular metalloprotease
IGFBP4	$+5.37$	IGF signaling modulator
SPHK1	$+4.87$	Sphingosine kinase; pro-fibrotic lipid signaling
IL32	$+4.87$	Pro-inflammatory cytokine
SAMD11	$+4.84$	Transcriptional regulator
SERPINE2	$+4.79$	Serine protease inhibitor; ECM remodeling
RUNX1	$+3.79$	Myeloid/fibroblast TF; differentiation driver

Top Downregulated Genes (Lost Homeostatic Program):

Gene	Log2FC	Functional Role
CYP4B1	-32.6	Xenobiotic metabolism; lung-specific detoxification
SAA1	-6.12	Acute phase protein
NKD1	-5.94	Wnt pathway negative regulator
HIF3A	-5.40	Hypoxia response modulator
CCN5	-3.55	Anti-fibrotic matricellular protein
ALDH1A1	-3.14	Retinoid metabolism; fibroblast quiescence
PDGFRA	-1.59	PDGF receptor; fibroblast identity marker

Step 3 Output: Functional Enrichment. ToppGene enrichment of the 56 upregulated genes identified the following top-ranked processes and pathways:

Term	p-value	Genes
Regulation of cell adhesion	2.47×10^{-9}	15/49
Regulation of cell migration	1.36×10^{-8}	16/49
Regulation of cell motility	2.98×10^{-8}	16/49
Positive regulation of cell adhesion	1.06×10^{-7}	—
Regulation of plasminogen activation	1.08×10^{-7}	—
Dissolution of fibrin clot (Reactome)	7.69×10^{-6}	—
Eicosanoid synthesis (WikiPathways)	4.06×10^{-5}	—
HIF-1 transcription factor pathway	4.96×10^{-5}	—
RUNX1-regulated myeloid differentiation	1.94×10^{-4}	—

The downregulated program was enriched for Phase I xenobiotic metabolism (CYP4B1, FMO2, INMT; $p = 6.01 \times 10^{-8}$), amino acid metabolism, and RECK pathway (MMP regulation), indicating loss of normal fibroblast detoxification capacity.

Key druggable targets identified by enrichment: kinases (MAPKAPK2, PIM3, NEK6), receptors (IL1R1, TNFRSF1A, LPAR1, AGTR1), and transcription factors (RUNX1, NFATC1/NFATC2, SOX4, FOXP1).

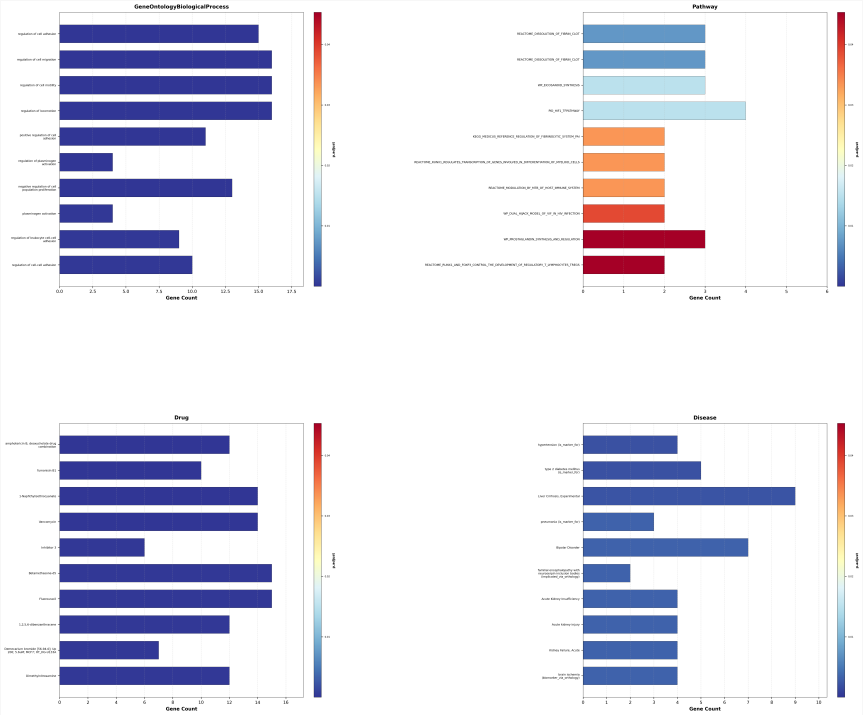

Figure E1: ToppGene functional enrichment of the top 56 upregulated IPF alveolar fibroblast genes. GO Biological Process terms highlight cell adhesion, migration, and HIF-1 signaling; Disease ontology maps to pulmonary fibrosis and related pathologies; Drug associations support kinase inhibitor candidates. All terms FDR-corrected, $q < 0.05$.

Step 4 Output: Spatial Context. Spatial transcriptomics (Xenium, single-cell resolution) revealed seven distinct fibroblast subtypes with segregated niches in unaffected lung tissue: Alveolar_FBs distributed throughout parenchyma, Subpleural_FBs at tissue edges, Myofibroblasts in focal clusters, and Activated_Fibrotic_FBs near remodeling zones. CellChat and Squidpy interaction analysis of Activated_Fibrotic_FBs (IPF vs. control) identified:

- **46 significant interactions** in More_Affected tissue; **33 unique to IPF**
- **Gained in IPF:** Interactions with KRT5^{-}/KRT17^{+} aberrant basaloid cells and capillary populations; monocyte–fibroblast crosstalk amplified
- **Lost in IPF:** Lymphatic and venous endothelial communication; NK cell interactions reduced
- **Top ligand–receptor pairs:** VEGFA–VEGFR2, EREG–EGFR, SPP1–ITGAV/ITGB1, SPP1–CD44, CCL2–ACKR1

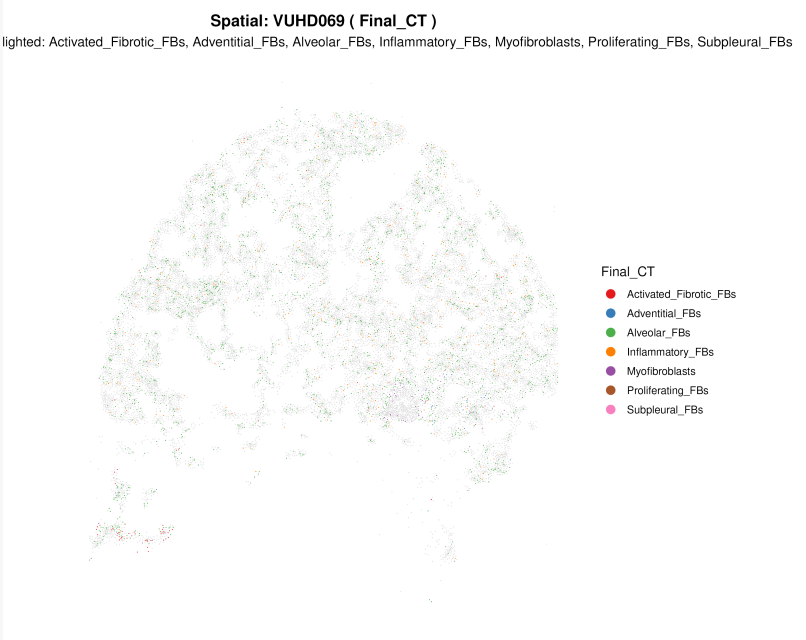

Figure E2: Spatial cell-type distribution in Unaffected IPF tissue (Xenium). UMAP overlay reveals fibroblast subpopulation geography across seven distinct niches.

Figure E3: CellChat communication profile for Activated Fibrotic Fibroblasts: More_Affected vs. Unaffected IPF regions. Key gained interactions include SPP1–CD44, VEGFA–VEGFR2, and CCL2–ACKR1 (46 total interactions in More_Affected, 33 unique to the fibrotic state).

Step 5 Output: iLINCS Drug Connectivity. The 3,563-gene alveolar fibroblast IPF signature was screened against **2,383 compounds** in the LINCSCP library (reversal mode). The top 15 candidates ranked by anticorrelation score:

Rank	Drug	Score	Targets	Mechanism
1	**Saracatinib**	$\mathbf{-0.451}$	ABL1, SRC	Src/Abl dual kinase inhibitor
2	TWS 119	-0.393	GSK3B	GSK-3β inhibitor (Wnt activator)
3	Tivozanib	-0.392	VEGFRs, PDGFRs	Multi-VEGFR/PDGFR TKI
4	LY2874455	-0.389	FGFR1–4	Pan-FGFR inhibitor
5	Linifanib	-0.386	CSF1R, FLTs, PDGFRs	Multi-RTK inhibitor
6	CI-1040	-0.385	MAP2K1/2	MEK1/2 inhibitor
7	Tozasertib	-0.385	ABL1, AURKs, JAK2	Aurora/Abl kinase inhibitor
8	Gefitinib	-0.383	EGFR	EGFR TKI
9	WH-4-025	-0.382	—	—
10	Tamatinib	-0.380	SYK	SYK kinase inhibitor
11	Fedratinib	-0.377	FLT3, JAK2	JAK2/FLT3 inhibitor
12	BX-795	-0.374	PDK1	PDK1 inhibitor
13	Canertinib	-0.370	EGFR, ERBB2/4	Pan-ErbB inhibitor
14	Pazopanib	-0.365	Multi-RTK	Broad-spectrum TKI
15	LDN 193189	-0.362	ACVR1, BMPR1A	BMP/ALK2 inhibitor

Mechanistic convergence: The top hits cluster around Src/Abl kinase, VEGFR/PDGFR signaling, FGFR, MEK/ERK, and JAK2, all pathways directly implicated in the enriched fibroblast program (cell adhesion, migration, ECM remodeling, HIF-1 signaling). Saracatinib was autonomously selected as the priority candidate for downstream DART analysis.

Steps 6–7 Output: DART Cell-Type Impact Analysis. The supervisor agent retrieved Saracatinib's own iLINCS compound perturbation signature and executed DART, correlating this signature against **475 lung cell-type/disease signatures** across HLCA, BPD_Sun, and HCA_fetal datasets.

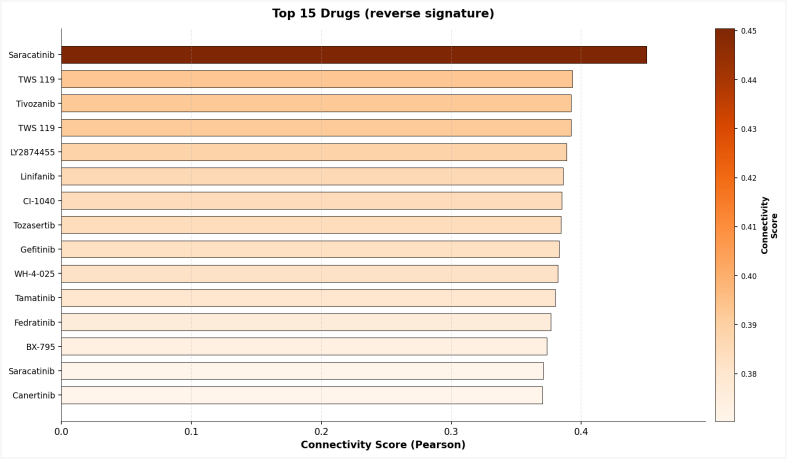

Figure E4: iLINCS drug connectivity screen for the IPF alveolar fibroblast signature (reversal mode, 2,383 compounds). Saracatinib (Src/Abl inhibitor) achieves the highest anticorrelation score ($r = -0.451$), followed by VEGFR/FGFR/MEK inhibitors, consistent with the enriched fibroblast transcriptional program.

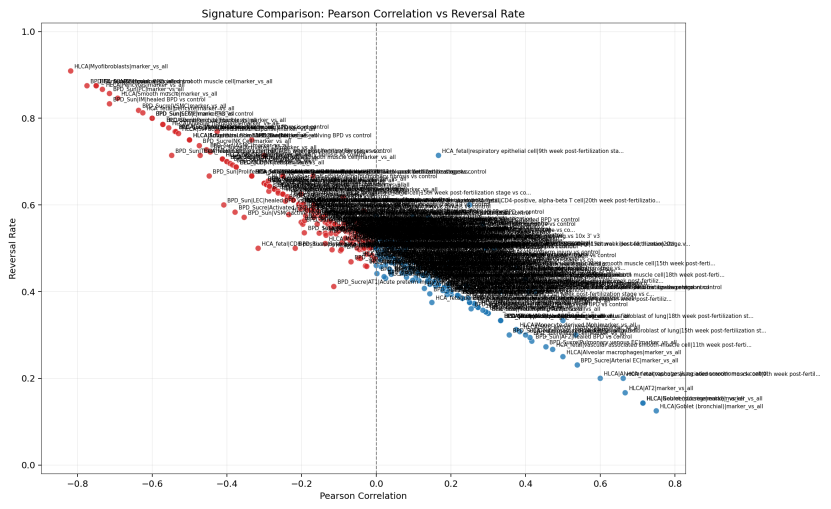

Figure E5: DART cell-type impact analysis for Saracatinib across 475 lung signatures. Negative Pearson r (blue) indicates therapeutic reversal of a signature; positive r (red) indicates mimicry. Saracatinib achieves strong reversal of myofibroblast ($r = -0.818$), smooth muscle, and pericyte marker programs, with a favorable safety profile in alveolar epithelial cells.

Strongest Reversal Signatures (On-Target Therapeutic Benefit):

Cell Type	Dataset	Comparison	r	RR	Key Reversed Genes
Myofibroblasts	HLCA	marker vs. all	$\mathbf{-0.818}$	**90.9%**	COL1A1, TPM1, MYL9, MYLK, MMP2, SERPINE1
CAP2 capillaries	BPD_Sun	BPD vs. ctrl	-0.775	87.5%	KAT6A, NCOA3, COL4A1, SOX4, MEF2C
AF1 fibroblasts	BPD_Sun	marker vs. all	-0.750	87.5%	MACF1, MYLK, MYO10, TBX2, GLI2
Vascular SMC	HCA_fetal	marker vs. all	-0.750	87.5%	MYL9, TPM1, COL4A1, COL1A1, MYLK
Pericytes	HLCA	marker vs. all	-0.733	86.7%	COL4A1, TBX2, TPM1, MYL9, MYLK
Smooth muscle	HLCA	marker vs. all	-0.692	84.6%	MYL9, TPM1, MYLK, COL4A1, COL1A1
Peribronchial FBs	HLCA	marker vs. all	-0.571	78.6%	COL1A1, MMP2, MYL9, S100A13, P4HA2
Alveolar FBs	HLCA	marker vs. all	-0.556	77.8%	MYLK, MMP2, TBX2, COL1A1, CSRP1

The reversal signature is dominated by contractile/ECM genes (COL1A1, MYL9, MYLK, TPM1, MMP2, SERPINE1), the core fibrotic effector program, across multiple mesenchymal and vascular cell types.

IPF-Specific Reversal Across Cell Types (HLCA disease signatures):

Cell Type	Pearson r	Reversal Rate
EC general capillary	-0.537	76.9%
Pre-TB secretory	-0.412	70.6%
EC arterial	-0.300	71.4%
Alveolar Mph proliferating	-0.289	65.5%
NK cells	-0.263	62.8%
CD8 T cells	-0.234	57.8%
Alveolar fibroblasts	-0.207	60.2%

Saracatinib's reversal effect extends beyond fibroblasts to vascular, immune, and epithelial compartments in IPF, consistent with Src kinase's broad role in cell signaling.

Step 8 Output: Safety Profile. DART identified positive correlations between Saracatinib's perturbation signature and normal cell-type identity programs, indicating potential off-target effects:

Cell Type	r	Concern
Goblet (bronchial)	$+0.750$	Mimics goblet identity genes (XBP1, SPDEF, MUC1); potential mucus secretion disruption
Goblet (subsegmental)	$+0.714$	Same mucus program concern
Neuroendocrine	$+0.714$	Mimics neuroendocrine markers (SNAP25, STMN1); monitor for neuroendocrine effects
AT2 cells	$+0.667$	Mimics AT2 identity (FGFR2); potential surfactant biology perturbation
Alveolar macrophages	$+0.500$	Mimics macrophage markers (PPARG, APOE); immune surveillance impact

Interpretation. Saracatinib exhibits strong reversal of fibrotic mesenchymal programs, particularly in myofibroblasts ($r = -0.818$, 90.9% reversal) and perivascular populations, with the effect extending across vascular and immune compartments, indicating system-level remodeling of the fibrotic niche. The positive correlations with goblet and AT2 cells indicate that Saracatinib's signature overlaps with normal epithelial identity programs; this does not necessarily indicate toxicity but warrants monitoring of mucociliary function and surfactant biology in preclinical models. Saracatinib demonstrated superior anti-fibrotic efficacy compared to nintedanib and pirfenidone in preclinical models Ahangari et al. (2022) and is currently being evaluated in a Phase 1b/2a clinical trial (STOP-IPF, NCT04598919), providing independent clinical validation of this computationally derived prediction.

This entire analysis was completed from a single user prompt in a fresh session with no prior conversation history, demonstrating LungChat's capacity for autonomous, zero-shot therapeutic discovery with complete provenance.

E.2 COPD: Fluticasone Propionate Discovery

Supplementary E2. COPD DART Case Study: Cell-Type-Resolved Zero-Shot Therapeutic Discovery

User Prompt (verbatim).

> *Identify therapeutic targets and drug repurposing candidates for COPD, separately considering each cell type. Among the top 10 hits for each cell type, which is the most promising therapeutic given prior clinical data for COPD in which cell type. For this drug, show me the top dysregulated genes in that cell population then evaluate its cell-type-specific impact across all lung populations in adults and in utero, to determine on-target efficacy and off-target safety risks.*

Supervisor Decomposition. LungChat's supervisor agent decomposed this single natural-language prompt into a four-stage pipeline with extensive parallelization across cell types:

1. **Per-cell-type DEG signatures:** Generate COPD vs. control DEGs independently for each of the 14 lung cell types in HLCA
2. **Per-cell-type iLINCS screening:** Run drug connectivity analysis on each cell type's signature in parallel, producing independent top-10 drug lists per population
3. **Clinical evidence cross-referencing:** Scan all top-10 lists for drugs with documented COPD clinical trial history; select the most clinically validated candidate
4. **DART cell-type impact:** For the selected drug, generate a volcano plot of DEGs in its primary cell type; then execute DART across adult (HLCA) and fetal (HCA_fetal) lung populations

Unlike the IPF case study, this workflow independently screens **14 cell types**, enabling cell-type-resolved drug discovery and clinical evidence cross-validation across populations.

Complete Execution Trace.

Step	Agent	Tool(s)	Operation
1	Supervisor	`decompose`	Parses query; identifies 14 cell types with COPD samples in HLCA; plans parallel iLINCS dispatch
2	Single-Cell	`generate_stats` (×14)	Differential expression for each of 14 cell types (COPD vs. control, HLCA). 8 cell types yield signatures with sufficient gene counts for iLINCS
3	Research	`drug_connectivity_tool` (×8, batched)	iLINCS drug connectivity analysis on each cell type's DEG signature independently (reversal mode, LINCSCP). Returns per-cell-type top-10 ranked drug lists
4	Research	`literature_search_tool`, `clinical_trials_search_tool`	Cross-references all top-10 hits against COPD clinical trial databases. Identifies Fluticasone Propionate (>200 trials, Phase 4 approved) in DC2 and Non-classical Monocyte lists
5	Single-Cell	`plot_differential_expression_volcano`	Volcano plot of 521 DEGs in DC2 cells (COPD vs. control)
6	Supervisor	`query_pseudobulk_expression_data_tool` (`drug_cell_type_impact`)	Executes DART: correlates Fluticasone Propionate iLINCS signature (978 genes) against 252 lung signatures across HLCA (adult) and HCA_fetal (in utero) datasets
7	Supervisor	Synthesis	Aggregates all outputs into integrated therapeutic report with on-target efficacy and fetal off-target safety profiling

Step 2–3 Output: Per-Cell-Type Drug Screening. Of 14 cell types with COPD DEG signatures, 8 yielded sufficient gene counts for iLINCS analysis. Independent screening produced cell-type-resolved drug rankings:

Cell Type	Top Hits (selected)	Clinically Relevant	Score Range
Alv. Macrophages	Manumycin-A, L-Sulforaphane	—	-0.31 to -0.37
AT2	Diphenylpyraline, Camptothecin	Sorafenib (#9), Bosutinib (#10)	-0.61 to -0.69
CD4 T cells	Emodic acid, Cerulenin	Tretinoin (#8)	-0.60 to -0.63
CD8 T cells	MD-041, Rita	Valdecoxib (#7)	-0.49 to -0.55
Classical Mono.	AZD-8330, 10-DEBC	Pimasertib (#7), Rimantadine (#8)	-0.40 to -0.45
DC2	Veliparib (#1), Piplartine	**Fluticasone Prop. (#6)**, Erlotinib (#9)	-0.62 to -0.68
NK cells	ML175, CAY10618	Momelotinib (#3)	-0.66 to -0.74
Non-cl. Mono.	Veliparib (#1), Piplartine	**Fluticasone Prop. (#6)**	-0.62 to -0.68

Step 4 Output: Clinical Evidence Cross-Referencing. Scanning all 80 candidate drugs (top 10 $\times$ 8 cell types) against COPD clinical trial databases:

Drug	Cell Types (rank)	COPD Clinical Evidence
Fluticasone Propionate	DC2 (#6), Non-cl. Mono. (#6)	**Approved.** >200 COPD trials. Phase 4 completed.
Tretinoin	CD4 T cells (#8)	Phase 2 only (emphysema subtype). Not approved.
Rimantadine	Classical Mono. (#8)	One trial (viral prophylaxis context only).
Others	Various	No COPD trials identified.

Fluticasone Propionate was selected as the priority candidate: the only drug in any top-10 list with Phase 4 COPD approval, emerging independently from both DC2 and Non-classical Monocyte signatures ($p = 3.4 \times 10^{-5}$).

Step 5 Output: DC2 Differential Expression (COPD vs. Control). Analysis of DC2 (type 2 dendritic) cells identified **521 DEGs** (COPD vs. control). The top dysregulated genes reveal a shift toward a mature, antigen-presenting phenotype with loss of innate alarm signaling:

Dir.	Gene	LogFC	FDR	Functional Role
$\uparrow$	CD207	$+2.49$	9.95×10^{-5}	DC maturation / antigen capture
$\uparrow$	PKP2	$+2.16$	6.22×10^{-7}	Desmosomal signaling
$\uparrow$	MT2A	$+1.77$	1.13×10^{-5}	Metallothionein, oxidative stress
$\uparrow$	F2RL2	$+1.69$	2.86×10^{-4}	Protease-activated receptor
$\uparrow$	CXCL9	$+1.64$	4.13×10^{-4}	T cell recruitment chemokine
$\downarrow$	S100A9	-1.74	4.42×10^{-5}	Alarmin / innate immune activation
$\downarrow$	CXCL8	-1.73	1.50×10^{-4}	Neutrophil recruitment (IL-8)
$\downarrow$	MMP14	-1.55	5.58×10^{-4}	Matrix metalloprotease
$\downarrow$	SIRPA	-1.25	2.64×10^{-6}	Phagocytosis checkpoint
$\downarrow$	TNFAIP8	-1.20	6.57×10^{-6}	NF-κB / apoptosis regulator
$\downarrow$	HIF1A	-1.05	1.28×10^{-5}	Hypoxia response
$\downarrow$	BCL2A1	-1.03	3.29×10^{-4}	Anti-apoptotic

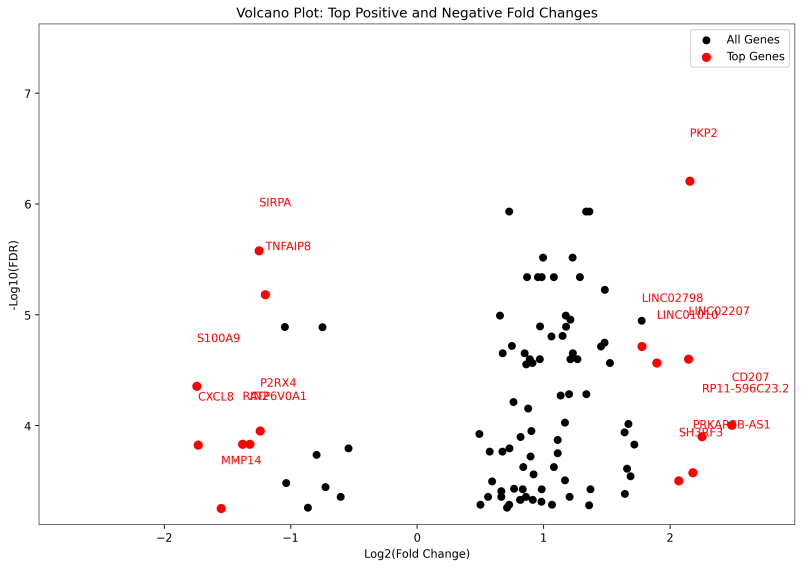

Figure E6: Volcano plot of differential gene expression in COPD vs. control DC2 cells (521 DEGs). Upregulated genes (CD207, CXCL9, PKP2) indicate enhanced antigen presentation; downregulated genes (S100A9, CXCL8, MMP14, HIF1A) indicate loss of innate alarm signaling and hypoxia response.

Step 6 Output: DART Cell-Type Impact Analysis. The Fluticasone Propionate iLINCS perturbation signature (**978 genes**) was correlated against **252 lung signatures** across HLCA (adult) and HCA_fetal (in utero) datasets. A 100-gene subset was used for impact analysis.

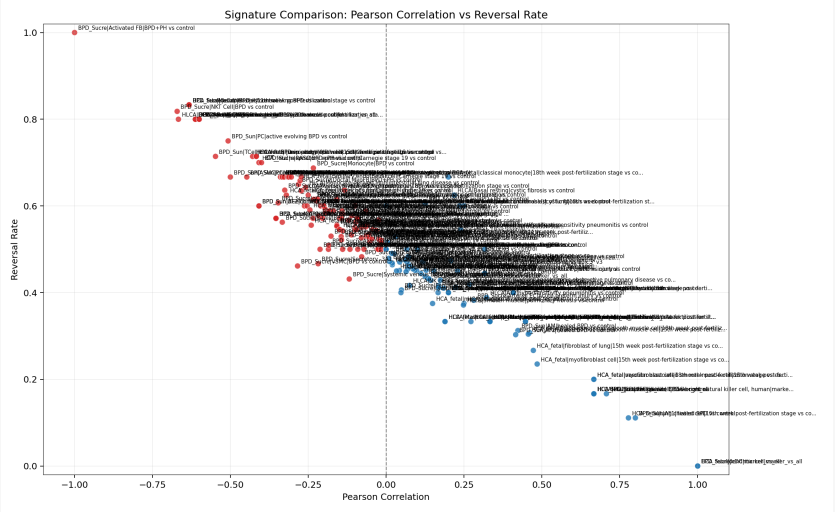

Figure E7: DART cell-type impact analysis for Fluticasone Propionate across 252 lung signatures (HLCA adult + HCA_fetal). Negative Pearson r (blue) indicates therapeutic reversal; positive r (red) indicates signature amplification. DC2 COPD signature shows strong reversal ($r = -0.667$, 80% RR), while fetal dendritic and ciliated cell developmental programs show positive correlation, indicating potential off-target effects during gestation.

On-Target Efficacy (Adult HLCA, COPD Signatures):

Cell Type	Pearson r	RR	Overlap	Key Reversed Genes
DC2 (COPD vs. ctrl)	$\mathbf{-0.667}$	**80%**	5	JUN↓, USP7↓, CCNH↓, MEF2C↓

The DC2 COPD signature is the only adult HLCA COPD signature with sufficient overlap, and it shows strong anticorrelation ($r = -0.667$, reversal rate 80%), confirming Fluticasone's predicted on-target activity in the cell type where it was identified.

Off-Target Safety (Fetal HCA_fetal, Developmental Programs):

Cell Type	r	RR	Overlap	Safety Interpretation
Dendritic cell (marker)	$+1.000$	0%	5	Drug matches fetal DC identity genes (PAK1, SYK, MEF2C); risk of disrupting fetal DC development
Ciliated cell (15 wk)	$+0.778$	11%	9	Drug amplifies ciliated cell developmental program; potential for altered airway epithelial maturation
Myofibroblast (18 wk)	$+0.667$	20%	5	Overlap with mesenchymal developmental genes (BMP4, CSRP1); possible off-target on lung structural development
Vascular SMC (15 wk)	$+0.667$	20%	5	Shared signature with vascular developmental program (COL1A1, MEF2C)

Interpretation. Fluticasone Propionate demonstrates strong reversal of the COPD-associated transcriptional program in DC2 dendritic cells ($r = -0.667$, 80% reversal rate), the primary cell type where it was computationally identified. The reversed genes (JUN, MEF2C, CCNH, USP7) are consistent with glucocorticoid receptor (NR3C1)-mediated suppression of inflammatory signaling, mechanistically linking the drug's known anti-inflammatory activity to the COPD DC2 disease program.

The fetal safety analysis reveals a clinically interpretable risk: Fluticasone's signature positively correlates with normal developmental programs in dendritic cells ($r = +1.0$), ciliated epithelial cells ($r = +0.778$), myofibroblasts ($r = +0.667$), and vascular smooth muscle ($r = +0.667$) at 15–18 weeks gestation. This transcriptomic signal aligns with Fluticasone's pregnancy Category C classification, providing cell-type-resolved molecular evidence for caution during fetal airway and mesenchymal development.

The finding that an approved, Phase 4-validated COPD corticosteroid is autonomously identified as the top candidate from cell-type-resolved computational screening, with mechanistically coherent on-target and off-target profiles, provides strong external validation of the DART method's capacity for clinically meaningful drug repurposing. The overlap counts are small (5–9 genes per comparison), so these correlations should be interpreted as directional signals rather than definitive predictions.

This entire analysis, including per-cell-type drug screening across 14 populations, clinical evidence cross-referencing, DC2 differential expression, and dual-dataset (adult + fetal) DART safety profiling, was completed from a single user prompt in a fresh session with no prior conversation history.

F Architecture Ablation Details

This appendix provides the full specification and detailed results for the architecture ablation described in Section 4.1.2.

F.1 Ablation Configuration Details

Table F1 details the four architecture configurations evaluated. The key design principle: flat-agent baselines preserve all domain knowledge from specialist prompts, ensuring that observed differences reflect architectural choices rather than information asymmetry.

Table F1: Ablation arm specifications. All configurations use the same foundation models, datasets, and evaluation rubrics.

Config	Agent(s)	Prompt	Tool Access	What It Isolates
MA+Sel (production)	5 agents (Supervisor + 4 specialists)	Specialized per agent	Supervisor routing + hybrid selector	Full system baseline
MA−Sel	5 agents (Supervisor + 4 specialists)	Specialized per agent	Supervisor routing, no selector	Selector contribution
Flat+Sel	1 agent	Merged flat prompt	All 51 tools + hybrid selector	Value of hierarchy when tool narrowing retained
Flat−Sel	1 agent	Merged flat prompt	All 51 tools, no selector	Hierarchy + tool narrowing combined

F.2 Context Engineering Rationale

The multi-agent architecture is motivated by a context-engineering principle: each specialist agent operates in a scoped context window containing only its domain-relevant tools and instructions, preventing cross-domain parameter leakage and instruction dilution. Token efficiency is a measurable consequence of this design (cumulative prompt tokens scale sub-linearly with task complexity), but the primary benefit is preserved reasoning quality under increasing tool and domain complexity.

Table F2 summarizes the specific context problems addressed by the hierarchical design.

Table F2: Context problems addressed by multi-agent hierarchy vs. flat architecture.

Problem	Flat Agent	Multi-Agent
Tool confusion	51 tool schemas compete for attention	Sub-agent sees ∼13 domain-relevant tools
Cross-domain contamination	HLCA metadata fields leak into spatial parameters	Each domain operates in isolated context
Instruction dilution	Combined prompts across all domains	Each sub-agent receives 13–29KB scoped instructions
Error containment	Bad tool call poisons entire reasoning chain	Sub-agent fails locally; supervisor retries or degrades
Scalability	Adding tools increases token overhead for ALL queries	New tools affect only their specialist agent
Tool attention bloat	Agent overly eager to invoke *some* tool	Routing layer abstracts tools from supervisor

The context-partitioning principles underlying our design align with several recent findings. Liu et al. Liu et al. (2024) demonstrated systematic "lost in the middle" degradation where LLM accuracy drops for mid-context information, a phenomenon exacerbated by monolithic context accumulation. Anthropic Anthropic (2025) identifies sub-agent delegation and context compaction as primary mitigations for context degradation in long-running agents. Zhang et al. Zhang et al. (2024) showed that multi-agent collaboration through scoped worker contexts outperforms single-agent systems by up to 10% on long-context tasks.

F.3 Full Single-Step Ablation Results

Table F3 provides the complete per-metric breakdown for all 50 single-step queries across 4 ablation modes (n=3 repetitions each, 600 total evaluations).

Table F3: Full single-step ablation results (50 queries × 3 reps × 4 modes = 600 evaluations). **Bold** indicates best per metric. Quality: higher is better. Efficiency: lower is better.

	Metric	MA+Sel	MA−Sel	Flat+Sel	Flat−Sel
Quality	LLM Judge	0.893 ± 0.31	0.927 ± 0.26	0.947 ± 0.23	**0.980** ± 0.14
	Tool Selection	0.920 ± 0.27	0.933 ± 0.25	0.927 ± 0.24	**0.980** ± 0.14
	Config Match	0.927 ± 0.26	0.940 ± 0.24	0.913 ± 0.28	**0.947** ± 0.23
	Items Coverage	0.891 ± 0.27	0.919 ± 0.23	**0.927** ± 0.22	0.924 ± 0.21
	Execution Success	0.973 ± 0.16	**0.980** ± 0.14	0.973 ± 0.16	0.973 ± 0.16
	Metadata Consistency	**1.000** ± 0.00	0.973 ± 0.16	0.973 ± 0.16	**1.000** ± 0.00
Efficiency	Mean Latency (s)	50.2	56.7	46.3	**39.5**
	Mean Tokens (K)	**197**	212	330	401
	Mean Tool Calls	**2.05**	2.18	3.79	3.37

Observations.

1. **Quality**: Flat−Sel achieves the highest LLM Judge score (0.980) and Tool Selection accuracy (0.980), but the absolute gap to the production configuration (MA+Sel) is $\leq$0.09 across all metrics.
2. **Token efficiency**: MA+Sel consumes $\sim$2$\times$ fewer tokens than Flat−Sel (197K vs. 401K), reflecting context partitioning across specialist agents.
3. **Latency**: Flat−Sel is fastest (39.5s) because it avoids supervisor$\rightarrow$sub-agent routing overhead, while MA+Sel adds $\sim$10s for delegation.
4. **Tool calls**: MA+Sel averages 2.05 tool calls vs. Flat−Sel's 3.37, because flat agents attempt more exploratory tool calls without hierarchical guidance.
5. **Metadata consistency**: MA+Sel achieves perfect metadata consistency (1.000), suggesting that scoped contexts improve parameter grounding for dataset-specific metadata fields.

F.4 Security and Grounded Abstention Results

Security (13 queries $\times$ 3 reps $\times$ 4 modes $= 156$ evaluations). All four architectures achieved 100% prevention of explicit tool execution and secret leakage under prompt injection, demonstrating that core supervisor prompt engineering and sandbox boundaries are robustly engineered regardless of tool configuration. MA+Sel showed slightly lower refusal coherence (82.1% vs. 92.3%) because the selector mechanism occasionally attempts to legitimately route complex adversarial language before recognizing hostile intent.

Grounded Abstention (17 queries $\times$ 3 reps $\times$ 4 modes $= 204$ evaluations). This benchmark provides the strongest architectural validation. When presented with out-of-scope queries, flat architectures (whose context windows expose all 51 tool schemas) become overly eager to invoke *some* tool (54.9% disallowed-tool blocking vs. 90.2% for MA+Sel). This +35% absolute improvement directly validates the context-engineering hypothesis: abstracting domain tools behind a routing layer forces the top-level agent to reason about intent before invoking capabilities, yielding superior grounded refusal. MA+Sel also produces fewer invalid output artifacts (86.3% clean vs. 70.6% for Flat−Sel), indicating that context isolation reduces spurious output generation.

F.5 Multi-Tool Orchestration Results

Based on the monotonic cost-quality ordering in the 4-arm single-step ablation (MA+Sel: best cost, Flat−Sel: best accuracy), the multi-tool and decomposition-stress benchmarks were evaluated on the two boundary configurations: MA+Sel (production) and Flat−Sel.

Quality (10 queries $\times$ 3 reps $\times$ 2 modes $= 60$ evaluations). Table F4 summarizes multi-tool orchestration results. Both architectures achieve identical strict pass rates (66.7%) and comparable mean rubric scores (MA+Sel: 3.10/4, Flat−Sel: 3.23/4). Per-question analysis reveals task-dependent architecture preference rather than a global advantage: MA+Sel wins on queries requiring gene regulatory network generation with literature search or multi-dataset cell frequency analysis (MT03, MT04, MT06, MT08), where the supervisor's decomposition into parallel sub-agent tasks produces more coherent synthesis. Flat−Sel wins on queries with broader tool orchestration (MT01, MT05, MT09, MT10), where direct access to all tools avoids routing overhead.

Table F4: Multi-tool orchestration ablation (10 queries × 3 reps × 2 modes). Quality: higher is better. Efficiency: lower is better. **Bold** indicates best per metric.

	Metric	MA+Sel	Flat−Sel
Quality	Execution Success	0.967 ± 0.18	**1.000** ± 0.00
	Mean Rubric Score (0–4)	3.10 ± 1.51	**3.23** ± 1.36
	Strict Pass Rate (%)	66.7	66.7
	Task Coverage (%)	80.0	**83.3**
	Trace Fidelity (%)	76.7	**80.0**
	Cross-Domain Synthesis (%)	73.3	**76.7**
Efficiency	Mean Latency (s)	303.5	**151.7**
	Mean Tokens (K)	**1,477**	1,540
	Mean Tool Calls	**12.0**	18.6
	Hard Failure Rate (%)	10.0	**6.7**

Efficiency. Flat−Sel is 2× faster (152s vs. 304s) due to the absence of supervisor→sub-agent delegation overhead. However, total token consumption is nearly identical (∼1.5M per query), indicating the latency difference arises from sequential API calls in the delegation chain, not from additional computation. Flat−Sel uses 55% more tool calls (18.6 vs. 12.0), compensating for the lack of hierarchical routing intelligence with broader exploratory tool invocation. The cost crossover relative to single-step queries is notable: on multi-tool queries, Flat−Sel is modestly cheaper, because each sub-agent invocation in the hierarchical architecture carries context setup overhead.

F.6 Decomposition Stress Results

Quality (10 queries × 3 reps × 2 modes = 60 evaluations). Table F5 summarizes decomposition-stress results. Both architectures achieve near-ceiling quality: >83% strict pass rates with zero hard failures and mean rubric scores >5.3/6. This is the highest-quality benchmark in the evaluation suite, demonstrating that LungChat's analytical quality on complex decomposition tasks comes from its domain engineering (curated tools, pre-computed warehouse, expert prompts) rather than from the multi-agent hierarchy.

Table F5: Decomposition stress ablation (10 queries × 3 reps × 2 modes). Quality: higher is better. Efficiency: lower is better. **Bold** indicates best per metric.

	Metric	MA+Sel	Flat−Sel
Quality	Mean Rubric Score (0–6)	**5.47** ± 1.45	5.33 ± 1.81
	Strict Pass Rate (%)	83.3	**86.7**
	Hard Failure Rate (%)	**0.0**	**0.0**
	Hypothesis Coverage (%)	**93.3**	86.7
	Evidence for Each Side (%)	**93.3**	86.7
	Epistemic Calibration (%)	90.0	90.0
	Evidence Diversity (%)	90.0	90.0
	Cross-Dataset Integration (%)	**96.7**	90.0
Eff.	Mean Latency (s)	570.2	**214.9**
	Mean Tokens (K)	**1,965**	1,983
	Mean Tool Calls	**21.9**	29.3

Mechanistic observations. MA+Sel shows a modest edge on criteria that specifically reward structured decomposition: hypothesis coverage (+6.6pp), evidence for each side (+6.6pp), and cross-dataset integration (+6.7pp). These criteria reward the supervisor's ability to dispatch independent analyses to specialist agents before synthesis. Flat−Sel matches or exceeds on mechanistic depth (90.0% vs. 83.3%), suggesting that direct tool access without routing intermediation can yield deeper single-branch reasoning. Flat−Sel is 2.7× faster (215s vs. 570s) with dramatically lower variance, making it more predictable for complex queries.

F.7 Combined Evaluation Summary

Table F6 provides a consolidated view across all five evaluation tracks (100 unique questions, 1,080 total evaluations). No single configuration dominates all dimensions.

Table F6: Combined evaluation summary across all five tracks. **Bold** indicates the better configuration per metric. Quality metrics are reported as percentages; token counts are in thousands.

Track (N)	Key Metric	MA+Sel	Flat−Sel	Winner
Single-Step (600)	LLM Judge (%)	89.3	**98.0**	Flat (−8.7pp)
Single-Step (600)	Mean Tokens (K)	**197**	401	MA (2× fewer)
Security (156)	No Leak / No Exec (%)	**100.0**	**100.0**	Tie
Security (156)	Mean Tokens (K)	**38**	89	MA (2.3× fewer)
Grounded Abstention (204)	No Disallowed Tools (%)	**90.2**	54.9	**MA (+35.3pp)**
Grounded Abstention (204)	Mean Tokens (K)	**131**	340	MA (2.6× fewer)
Multi-Tool (60)	Mean Rubric (0–4)	3.10	**3.23**	≈Tie (−0.13)
Multi-Tool (60)	Mean Tokens (K)	**1,477**	1,540	≈Tie
Decomp. Stress (60)	Mean Rubric (0–6)	**5.47**	5.33	≈Tie (+0.14)
Decomp. Stress (60)	Mean Tokens (K)	**1,965**	1,983	≈Tie

The production multi-agent architecture (MA+Sel) achieves the best profile across the Pareto frontier: 89% single-step accuracy (vs. 98% for Flat−Sel), 90% disallowed-tool blocking (vs. 55% for Flat−Sel), comparable multi-tool and decomposition-stress reasoning, and 2× token efficiency on the dominant single-step query class. The flat-agent baseline demonstrates that LungChat's domain knowledge, not its architecture, is the primary driver of analytical quality, while the multi-agent hierarchy provides the safety, cost, and scalability properties required for production deployment.

F.8 Discussion and Limitations

Across all five evaluation tracks (100 questions, 1,080 evaluations), analytical quality is architecture-neutral: both multi-agent and flat-agent configurations achieve comparable accuracy on single-step, multi-tool, and decomposition-stress queries. The multi-agent hierarchy's measurable contribution is on dimensions that accuracy-only benchmarks do not capture: a +35pp improvement in grounded abstention (preventing the system from invoking biological tools on out-of-scope queries), $2\times$ token-cost reduction on the dominant single-step query class, and structural guarantees for error containment and extensibility.

The current evaluation focuses on single-turn queries. While the stateless sub-agent design provides structural advantages for multi-turn session continuity, where each sub-agent invocation operates in a fresh context independent of prior turns, we have not empirically evaluated this property across long sessions. At `chat.lungmap.net`, researchers routinely conduct sessions spanning 10–20 queries; in a monolithic architecture, accumulated context would approach 4M+ tokens after 10 turns, while the multi-agent supervisor carries compact summaries ($\sim$100K tokens) with each sub-agent receiving a fresh context. Quantifying quality degradation over extended sessions is an important direction for future work.

A natural next step is to distill the current LLM-based tool reranker into a lightweight learned reranker trained on successful traces, while retaining LLM fallback for low-confidence or out-of-distribution cases.