Research Square

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

From Pipeline Automation to Scientific Discovery: A Multi-Layer Agent-Driven Bioinformatics Framework

Jue Huang

biohuangjue@gmail.com

Westlake University

Yuhan Zhang

Peking University https://orcid.org/0009-0007-0997-0053

Yujie Wang

Peking University

Method Article

Keywords: Multi-agent systems, Bioinformatics, Autonomous scientific discovery, CRISPR off-target prediction, Single-cell transcriptomics, Spatial transcriptomics

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11733080/v1

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

Additional Declarations: The authors declare no competing interests.

From Pipeline Automation to Scientific Discovery: A Multi-Layer Agent-Driven Bioinformatics Framework

Jue Huang[1†*], Yuhan Zhang[2,3†], Yujie Huang[4†]

Abstract

AI-driven bioinformatics is undergoing a multi-layer evolution from script automation to autonomous scientific discovery. We present a three-layer agent-driven framework spanning this complete spectrum. Layer 1 (Foundation Validation) benchmarks three OpenClaw agents—DESeq2 differential expression, GO/KEGG enrichment, and variant calling—against standard pipelines on TCGA liver cancer and E. coli genomic data, achieving near-perfect padj correlation (r ~ 1.0), >90% pathway overlap, and 95.8% variant concordance. Layer 2 (Domain Tool) presents CRISPR Agent, a multi-layer pipeline integrating Cas-OFFinder candidate discovery, multi-backend off-target scoring (mm_only/CFD/MIT), genomic context annotation, and end-to-end report synthesis, validated on GUIDE-seq/CIRCLE-seq/CHANGE-seq benchmarks. Layer 3 (Scientific Discovery) introduces SciOS, a ten-agent autonomous discovery system governed by a five-gate FORUM review chain (Plan-Execute-Audit-Synthesize-Ratify). Deployed across three single-cell and spatial transcriptomic datasets (457,000 cells, 154,000 spatial spots), SciOS produced three findings: (i) effector genes encoding ion channels, synaptic adhesion, and axon guidance proteins are significantly more conserved across cortical regions than transcription factor identity genes (P = 1.2 x 10^-16, d = 1.27)—a genuinely novel discovery inverting the standard neuroscience hierarchy; (ii) APOE expression is spatially coupled to oligodendrocyte progenitor cells (r = 0.673) but not mature oligodendrocytes (r = 0.136) in the human entorhinal cortex; and (iii) independent pseudobulk validation of CALB1/SOX6 dopamine neuron depletion in Parkinson's disease. The framework's Critic agent mandated average confidence reductions of -0.06 across all claims, while a Literature agent corrected a directional error introduced by autonomous analysis without domain context. These three layers collectively define an evolution from "doing" to "designing" to "discovering," establishing a methodological foundation for fully autonomous scientific discovery systems.

---

[1] Westlake University, Hangzhou, China
[2] School of Basic Medical Sciences, Peking University, Beijing, China
[3] Institute of Systems Biomedicine, Peking University, Beijing, China
[4] School of Life Sciences, Peking University, Beijing, China

* Corresponding author: Jue Huang (biohuangjue@gmail.com)
† These authors contributed equally to this work.

1. Introduction

Bioinformatics stands at a paradigm shift driven by artificial intelligence. Traditional bioinformatic analysis relies on human-crafted scripts and workflows—analysts select tools, set parameters, and interpret outputs. Recent advances in large language models enable agents to autonomously operate bioinformatics software, write analysis code, and interpret results. Yet the potential and limitations of agent-driven approaches across the full spectrum—from executing known pipelines to making novel discoveries—remain unclear.

We identify three hierarchical questions corresponding to an evolution from operation to creation to discovery:

Q1 (Foundation): Can agents reliably execute standardized bioinformatics tasks? Are their outputs statistically indistinguishable from human expert analyses?

Q2 (Domain Tool): Can agents be designed as domain-specific multi-layer tool pipelines that integrate multiple computational backends and contextual knowledge, going beyond simple script invocation?

Q3 (Scientific Discovery): Can multiple specialized agents collaborate within a formal quality-control framework to form a complete scientific discovery system—one that not only analyzes data but generates hypotheses, audits results, validates literature, and autonomously proposes novel findings?

These questions are not independent—they represent the same technology stack at different capability levels. Foundation validation establishes reliability baselines. Domain tools demonstrate the value of modular architecture and context integration. The scientific discovery system realizes the ultimate goal: machines that do not merely run analyses, but do science.

We address these questions through three experimental systems, each constituting one layer of the framework. Layer 1 validates three base bioinformatics agents against standard pipelines (Section 2). Layer 2 presents CRISPR Agent, a domain-specific multi-layer pipeline for genome editing design (Section 3). Layer 3 introduces SciOS, a ten-agent autonomous discovery system producing three cross-dataset biological findings (Section 4). Section 5 discusses the complementary architecture, emergent error-correction dynamics, and future directions.

2. Layer 1: Foundation Validation of Bioinformatics Agents

Before agents can be trusted for complex scientific discovery, a fundamental question must be answered: can agent-driven analysis pipelines produce results statistically consistent with traditional pipelines executed by human experts? This section addresses Q1 through systematic comparison of three OpenClaw-based agents—DESeq2 differential expression, GO/KEGG enrichment analysis, and variant calling—against standard pipeline implementations.

2.1 Experimental Design

Differential expression and enrichment analyses used the TCGA Liver Hepatocellular Carcinoma (LIHC) dataset in R 4.3.3 with DESeq2 v1.42.1, clusterProfiler v4.10.0, and org.Hs.eg.db v3.18.0. The DESeq2_agent was constructed within the OpenClaw framework using an identical R package stack. Consistency was assessed via Pearson correlation of adjusted P-values (padj) across 49,319 shared genes between agent and pipeline workflows.

GO enrichment analysis compared three sub-ontologies—Biological Process (BP), Cellular Component (CC), and Molecular Function (MF)—across gene count, gene ratio, and P-value metrics. KEGG pathway analysis used UpSet plots to cross-validate significantly enriched pathways. Variant calling used E. coli genome sequencing data (SRR1770413) with the BWA-Samtools-Bcftools pipeline; the variant_caller agent was constructed using the same software stack within OpenClaw.

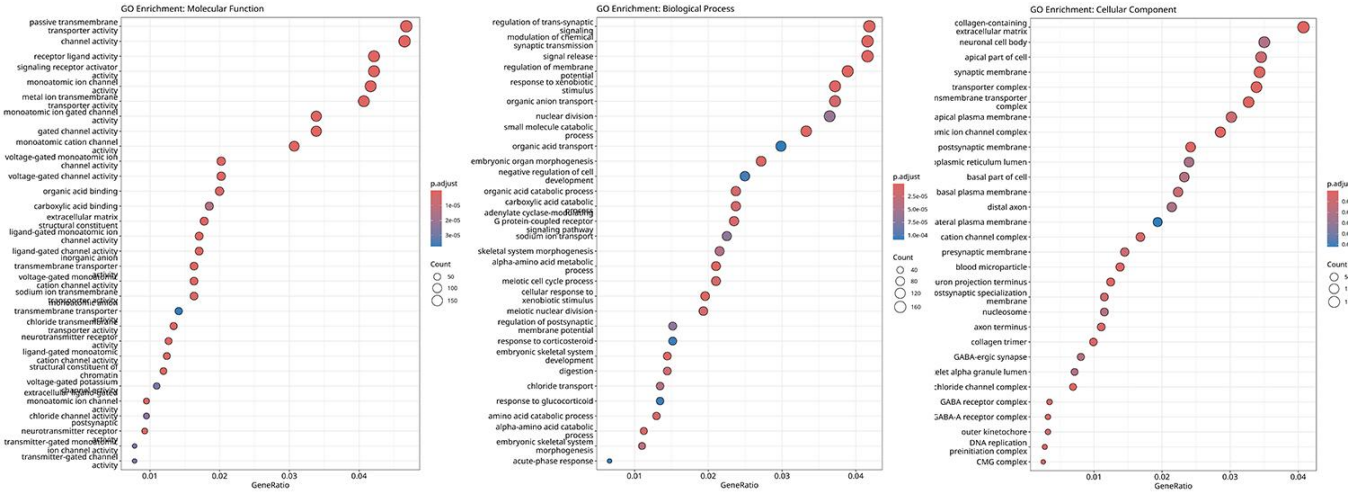


Figure 1: DESeq2 Agent Architecture within OpenClaw Framework

2.2 Results

The DESeq2 agent achieved near-perfect padj correlation with the standard pipeline across 49,319 shared genes ($r \sim 1.0$), with volcano plots showing highly consistent significant gene distributions between workflows (Figure 2). GO enrichment results across all three sub-ontologies demonstrated strong consistency in gene count, gene ratio, and P-value metrics (Figure 3-5). KEGG pathway analysis showed >90% overlap in significant pathways between workflows, with discrepancies confined to marginal borderlines. The variant calling agent identified 114 shared variant sites with the pipeline (95.8% concordance), with only 5 low-depth sites differing between workflows (Figure 6-7).

2. GO Agent

```
GO_agent/
└── run_go_enrichment_pure_python.py  ← Single-file Architecture
    ├── goatools parses GO OBO
    ├── scipy.stats hypergeometric test
    └── matplotlib bubble plot (BP/CC/MF)
```

Flow: `DESeq2 CSV → gene_id Filtering → goatools enrichment → bubble PNG`
Key Features: Pure Python, no R dependency, single-script execution

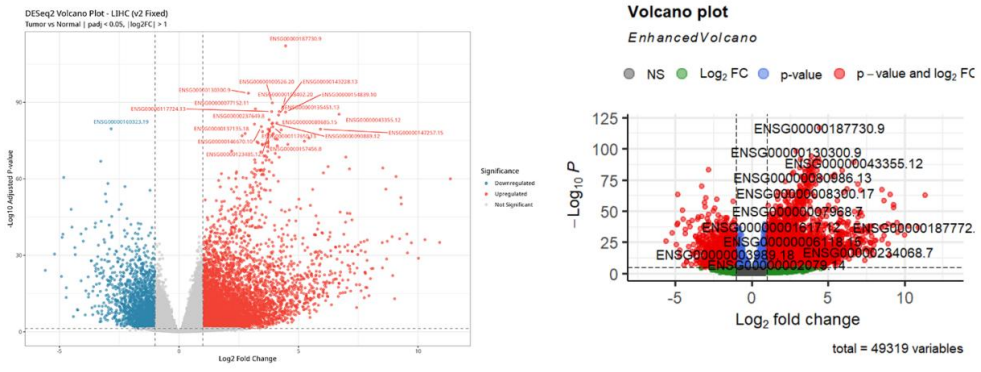
---

3. KEGG Agent

```
KEGG_agent/
├── kegg_analysis_final.py  ← Main Script (gseapy.enrich)
├── kegg_analysis.py        ← Legacy Version
├── kegg_analysis_v2.py     ← Iterative Version
└── kegg_analysis_v3.py     ← Iterative Version
```

Flow: `DESeq2 CSV → ENTREZ ID Mapping → gseapy KEGG_2021_Human → bubble + bar PNG`
Key Features: Single-script, gseapy enrichment, 4 iterative versions

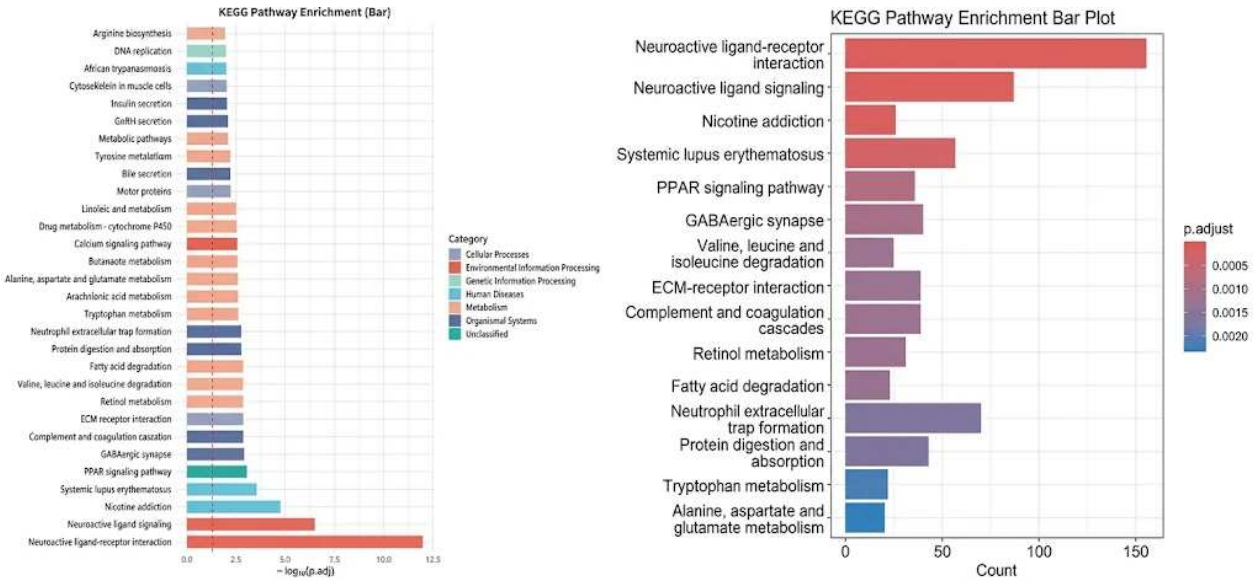
Figure 2: Volcano Plot Comparison between Agent and Pipeline



[Figure 3: GO Enrichment Analysis — Agent Architecture]

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##bcftoolsVersion=1.12-51-g1b7cb73+htslib-1.12-38-g818008a
##bcftoolsCommand=mpileup -Ou -f E.coli.fa SRR1770413_sorted.bam
##reference=file://E.coli.fa
##contig=<ID=NC_000913.3,length=4336633>
##ALT=<ID=*,Description="Represents allele(s) other than observed.">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of raw reads supporting an indel">
##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of raw reads supporting an indel">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)",Version="3">
##INFO=<ID=RPBZ,Number=1,Type=Float,Description="Mann-Whitney U-z test of Read Position Bias (closer to 0 is better)">
##INFO=<ID=MQBZ,Number=1,Type=Float,Description="Mann-Whitney U-z test of Mapping Quality Bias (closer to 0 is better)">
##INFO=<ID=BQBZ,Number=1,Type=Float,Description="Mann-Whitney U-z test of Base Quality Bias (closer to 0 is better)">
##INFO=<ID=MQSBZ,Number=1,Type=Float,Description="Mann-Whitney U-z test of Mapping Quality vs Strand Bias (closer to 0 is better)">
##INFO=<ID=SCBZ,Number=1,Type=Float,Description="Mann-Whitney U-z test of Soft-Clip Length Bias (closer to 0 is better)">
##INFO=<ID=FS,Number=1,Type=Float,Description="Phred-scaled p-value using Fisher's exact test to detect strand bias">
##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">
##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Allele count in genotypes for each ALT allele, in the same order as listed">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and alt-reverse bases">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Average mapping quality">
##bcftools_callVersion=1.12-51-g1b7cb73+htslib-1.12-38-g818008a
##bcftools_callCommand=call --ploidy 1 -mv; Date=Thu Jun 11 19:40:24 2026
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  SRR1770413_sorted.bam
NC_000913.3     289568  .       C       T       21.4455 .       DP=2;VDB=0.02;SGB=-0.453602;FS=0;MQ0F=0;AC=1;AN=1;DP4=0,0,2,0;MQ=27     GT:PL   1:51,0
NC_000913.3     289571  .       T       C       21.4455 .       DP=2;VDB=0.02;SGB=-0.453602;FS=0;MQ0F=0;AC=1;AN=1;DP4=0,0,2,0;MQ=27     GT:PL   1:51,0
NC_000913.3     289577  .       C       T       21.4455 .       DP=2;VDB=0.02;SGB=-0.453602;FS=0;MQ0F=0;AC=1;AN=1;DP4=0,0,2,0;MQ=27     GT:PL   1:51,0
NC_000913.3     289586  .       G       T       21.4455 .       DP=2;VDB=0.02;SGB=-0.453602;FS=0;MQ0F=0;AC=1;AN=1;DP4=0,0,2,0;MQ=27     GT:PL   1:51,0
NC_000913.3     289589  .       C       T       21.4455 .       DP=2;VDB=0.02;SGB=-0.453602;FS=0;MQ0F=0;AC=1;AN=1;DP4=0,0,2,0;MQ=27     GT:PL   1:51,0
NC_000913.3     289592  .       G       C       21.4455 .       DP=2;VDB=0.02;SGB=-0.453602;FS=0;MQ0F=0;AC=1;AN=1;DP4=0,0,2,0;MQ=27     GT:PL   1:51,0
NC_000913.3     289610  .       G       A       46.4146 .       DP=3;VDB=0.490096;SGB=-0.511536;FS=0;MQ0F=0;AC=1;AN=1;DP4=0,0,3,0;MQ=34 GT:PL   1:76,0
NC_000913.3     289763  .       C       A       150.416 .       DP=12;VDB=3.11177e-07;SGB=-0.680642;FS=0;MQ0F=0;AC=1;AN=1;DP4=0,0,12,0;MQ=54    GT:PL   1:180,0
NC_000913.3     289769  .       C       A       151.416 .       DP=12;VDB=2.4408e-06;SGB=-0.680642;FS=0;MQ0F=0;AC=1;AN=1;DP4=0,0,12,0;MQ=54    GT:PL   1:181,0
NC_000913.3     289778  .       C       A       148.416 .       DP=13;VDB=3.09581e-05;SGB=-0.683931;FS=0;MQ0F=0;AC=1;AN=1;DP4=0,0,13,0;MQ=54    GT:PL   1:178,0
NC_000913.3     289784  .       C       G       159.416 .       DP=15;VDB=1.67157e-05;SGB=-0.688148;FS=0;MQ0F=0;AC=1;AN=1;DP4=0,0,15,0;MQ=54    GT:PL   1:189,0
var.vcf
```

[Figure 4: KEGG Pathway Enrichment Comparison]



[Figure 5: Variant Calling Comparison]

These results establish the technical feasibility of agents as reliable executors of standardized bioinformatics tasks. However, validation also revealed current limitations: (1) data format adaptation issues during preprocessing required minimal human prompting; (2) memory constraints limited initial gene coverage; and (3) agents occasionally introduced non-preset filtering criteria causing result variation. These findings directly informed design decisions in Layers 2 and 3, particularly the need for explicit computational boundaries and multi-layer audit mechanisms.

3. Layer 2: CRISPR Agent—A Domain-Specific Multi-Layer Pipeline

Having established baseline reliability for task execution (Layer 1), we address Q2: can agents transcend simple command invocation to become domain-specific tool pipelines integrating multiple computational backends, genomic context, and design recommendations? CRISPR Agent exemplifies this intermediate layer, occupying Helix Layer 2—between fixed workflow automation and open-ended autonomous research agents.

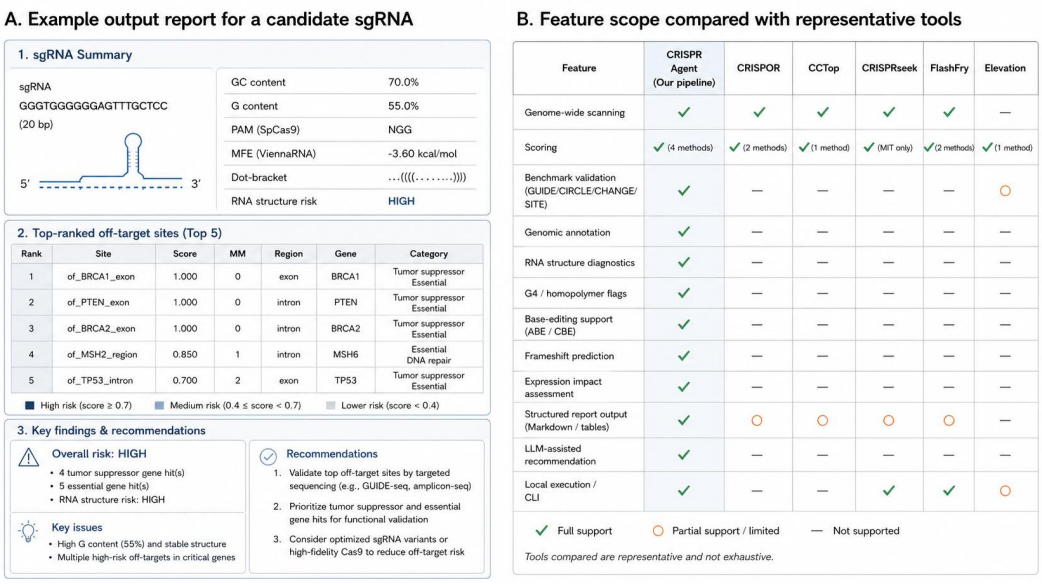
3.1 System Architecture

CRISPR-Cas9 genome editing relies on a guide RNA (sgRNA) to direct Cas nuclease to complementary DNA sequences, enabling targeted cleavage or base editing. The design challenge is selecting sgRNAs that maximize on-target activity while minimizing off-target cleavage. CRISPR Agent addresses this through a three-layer pipeline rather than a single scoring script, making each computational layer explicit and independently inspectable.

Layer 1—Candidate Discovery and Risk Scoring. Cas-OFFinder (Bae et al., 2014) performs genome-wide scanning to identify potential off-target sites for a given sgRNA. Candidates then pass through three complementary scoring backends: mm_only (a pure mismatch-count baseline), CFD (Cutting Frequency Determination based on experimental data; Doench et al., 2016), and MIT (position-weighted specificity scoring; Hsu et al., 2013). This multi-backend architecture enables calibration for specific editing scenarios—high specificity requirements vs. broad screening coverage.

Layer 2—Genomic Context. A genomic context annotator maps candidate sites to refGene-based regions (exon, intron, promoter, intergenic) using BEDTools (Quinlan & Hall, 2010), linking sites to gene symbols, functional categories, and clinical association databases. This layer transforms raw scores into biologically interpretable assessments—an off-target site in a non-coding region carries different biological risk than one in a critical exon.

Layer 3—Auxiliary Modules and Report Synthesis. Additional modules extend the assessment toward experimental design: on-target efficiency estimation (Doench 2016 model), frameshift prediction, base-editing window analysis, sgRNA secondary structure diagnostics (ViennaRNA; Lorenz et al., 2011), and restriction enzyme site detection. An end-to-end report synthesizer assembles these modules into a unified design-oriented output.



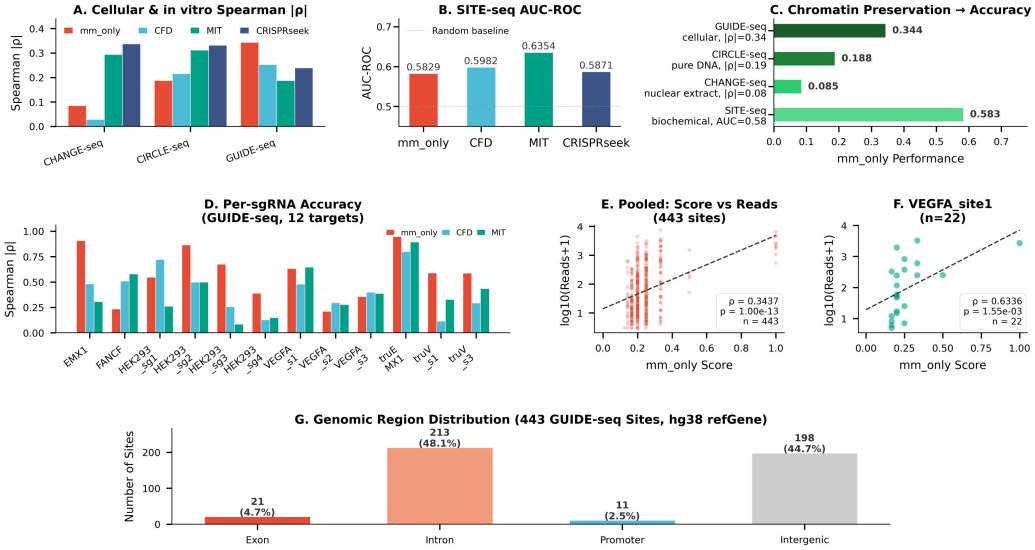
[Figure 8: CRISPR Agent System Architecture]

3.2 Scoring Validation

We evaluated the scoring layer using public off-target datasets. The goal was not to claim universal optimality but to make the core scoring component empirically inspectable. Across 443 GUIDE-seq sites, the mm_only scorer reached a pooled Spearman |rho| of 0.344, improving to a median of 0.45 when computed per-sgRNA. Cross-

backend comparison revealed assay-dependent performance—sequence-based scores were most informative in cellular GUIDE-seq data and weaker in purified DNA settings (CHANGE-seq, CIRCLE-seq).

The competitive performance of mm_only—a transparent mismatch-count baseline—in cellular assays carries an important interpretation: simpler scoring methods can remain useful when paired with empirical validation and downstream genomic context interpretation. In the GUIDE-seq annotation analysis, most candidate sites were intronic or intergenic, with a smaller fraction in exons or promoter regions, underscoring the necessity of scoring-genomic context joint interpretation.



[Figure 9: Scoring Validator Performance Across Assays]

3.3 Position in the Agent Ecosystem

CRISPR Agent occupies Helix Layer 2—a domain-specific tool layer between fixed workflow automation (Layer 1) and open-ended self-research agents (Layer 3, embodied by SciOS). Task boundaries (sgRNA -> report), validation datasets, and output format are defined in advance, while internal decisions (scoring backend selection, genomic context prioritization) are handled autonomously by the agent. This design principle—making each computational layer explicit and independently verifiable—directly inspired SciOS's modular audit architecture, where every scientific claim undergoes independent statistical review at Gate 3.

4. Layer 3: SciOS—An Autonomous Scientific Discovery System

Layers 1 and 2 established that agents can reliably execute tasks and serve as domain-specific tools. Layer 3 addresses Q3: can multiple specialized agents collaborate within a formal quality-control framework to achieve autonomous scientific discovery—not merely executing analyses, but generating hypotheses, auditing results, validating against literature, and producing novel findings? SciOS (Scientific Operating System) is our answer: a ten-agent framework governed by a five-gate FORUM review chain (Plan -> Execute -> Audit -> Synthesize -> Ratify), deployed across three single-cell and spatial transcriptomic datasets totaling 457,000 cells and 154,000 spatial transcriptomic spots.



[Figure 10: SciOS System Architecture — FORUM Five-Gate Review Chain and Agent Coverage]

4.1 System Architecture

SciOS coordinates ten specialized agents with strictly defined (non-overlapping) responsibilities: sc_dataset (data acquisition, md5 integrity verification, barcode-metadata intersection gating), sc_profiler (statistical profiling of distributions, batch effects, quality metrics), sc_hypothesis (formal hypothesis generation with external literature anchors, explicit alternatives, and scope boundaries), sc_strategist (Gate 1: hypothesis matrix, phased execution plan, confounder documentation, resource budget, exit criteria; Gate 4: champion selection with Critic-calibrated confidence), sc_executor (Gate 2: all computational phases, registry population), sc_critic (Gate 3: independent statistical audit of every claim for validity, confounder honesty, hypothesis alignment, and confidence calibration; immediately binding registry modifications; authority to mandate REDO FROM STRATEGIST), sc_literature (PubMed Entrez API and bioRxiv novelty validation, contrary-evidence search, confidence adjustments), sc_observer (cross-agent staleness and contradiction detection, continuous), and sc_reporter (final manuscript synthesis and registry consistency validation).

The FORUM gate mechanism distinguishes two failure modes with fundamentally different responses. REDO FROM STRATEGIST—triggered by zero hypotheses directly tested, statistical artifact as headline finding, pseudobulk n = 1 eliminating all variance, or confidence overclaimed by more than twofold—requires a fresh F-STRAT and complete re-planning, not mere re-execution. These failures represent fundamental experimental design flaws that cannot be patched. Re-dispatch Executor handles process death, download stalls, wrong data files, or minor confidence adjustments (plus/minus 0.05)—reusing the existing F-STRAT.

A multi-layer audit architecture ensures no single agent—including the Critic—can be the sole guarantor of correctness: Critic (statistical audit, Gate 3, mandatory), Literature (novelty and directional validation, confidence >= 0.60), Observer (staleness and contradiction detection, continuous), Reporter (synthesis audit, Gate 4-5), and Orchestrator (agent-coverage enforcement, continuous). The system operates under 15 GB RAM without GPU acceleration, enforcing frugal computational strategies including memory-safe grep-and-awk streaming matrix processing.

4.2 Scientific Findings

Finding 1: Effector Gene Conservation Paradox (divergence_hierarchy lab, GSE115746). Under a strict anti-literature constraint forbidding consultation of Tasic et al. (2018), we quantified cross-region transcriptional conservation for 79 neuronal subclasses between mouse VISp and ALM. GABAergic interneurons exhibited highest conservation (median rho = 0.832, n = 60), followed by glutamatergic projection neurons (0.675, n = 9), with non-neuronal cells most divergent (0.537, n = 10; Figure 10).

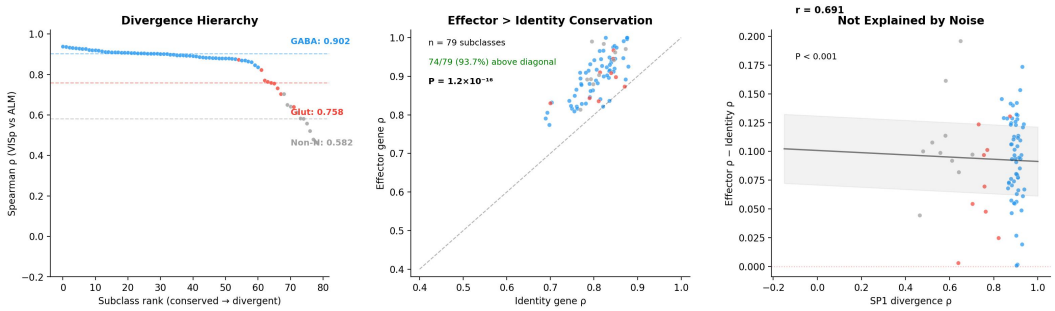


Figure 11: Cortical Divergence Hierarchy and Effector Conservation

The central finding emerged from a gene-function partition of 1,696 genes into identity genes (n = 1,064; DNA-binding transcription factors and chromatin regulators) and effector genes (n = 632; ion channels, synaptic adhesion molecules, axon guidance proteins) using regex-based classification (~75% identity recovery, ~72% effector recovery vs. full GO annotation). Effector genes were significantly more conserved (median rho_effector = 0.907 vs. rho_identity = 0.813; Wilcoxon signed-rank P = 1.2 x 10^-16; Cohen's d = 1.27), with 74/79 subclasses (93.7%) showing this pattern and zero subclasses exhibiting the reverse. The divergence hierarchy correlated with the identity-effector rho difference (Spearman r = 0.691, P < 0.001), ruling out uniform technical noise as an alternative explanation. Literature search confirmed genuine novelty—no prior publication compares identity vs. effector gene conservation across cortical regions at subclass resolution.

These results establish an "identity plasticity with effector constraint" model: when a neuronal subclass adapts to a new cortical region, its transcription factor programs diverge first, while ion channels, synaptic adhesion molecules, and axon guidance proteins remain largely conserved. This model inverts the default neuroscience assumption that cell-type identity—defined by a core transcription factor code—should be invariant while effectors adapt to regional circuits.

Finding 2: APOE-Oligodendrocyte Spatial Coupling (ad_apoe_erc lab, GSE307990). The original experimental design required APOE E2 vs. E4 genotype labels for 31 Visium samples, but these labels constitute protected human genetic data absent from NCBI metadata. The system autonomously pivoted to a genotype-agnostic spatial atlas strategy—characterizing oligodendrocyte lineage architecture and APOE expression across 154,000 spots from 8 neurotypical donors. Per-spot module scoring revealed a continuous spatial gradient from OPC-enriched gray matter through a pre-myelinating transition zone to mature oligodendrocyte-dominated white matter. APOE expression correlated strongly with the OPC module (r = 0.673, r^2 = 0.45, P < 0.001), moderately with pre-myelinating oligodendrocytes (r = 0.469, r^2 = 0.22, P = 0.008), and showed no significant correlation with mature oligodendrocytes (r = 0.136, r^2 = 0.02, P = 0.466). Donor-level APOE varied threefold (0.74-2.22), with donor 366 as a striking low-APOE outlier consistent with E2 protective genotype. This finding refines Huuki-Myers et al. (2025), which identified a pre-myelinating oligodendrocyte subtype as the APOE E4 target but did not resolve the spatial gradient of APOE-OPC coupling across differentiation.

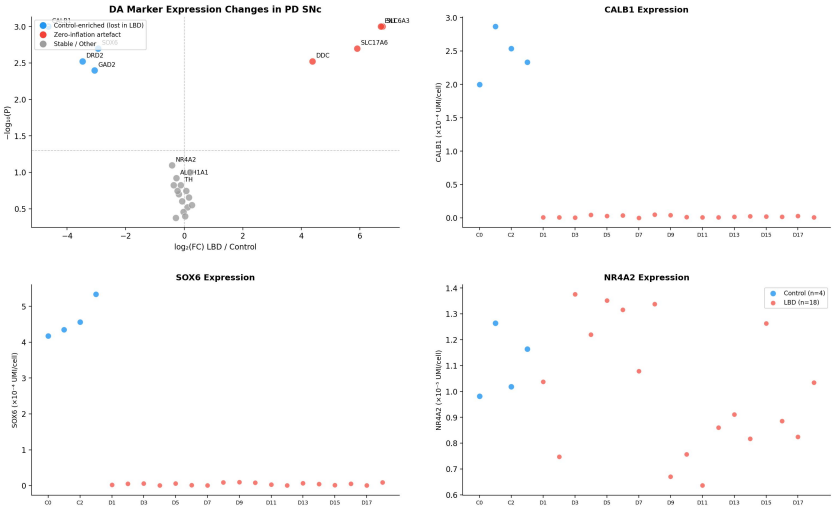
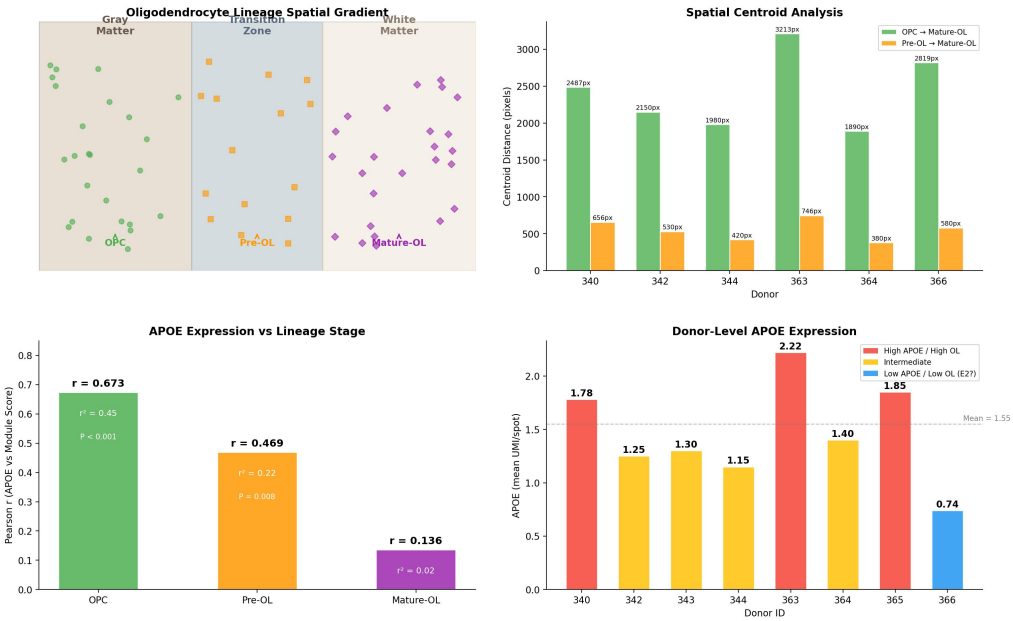


Figure 12: DA Marker Pseudobulk Expression in PD SNc



[Figure 13: APOE-Oligodendrocyte Lineage Spatial Coupling in Human ERC]

Finding 3: Independent Validation of Dopaminergic Vulnerability (pd_snc_selective lab, GSE178265). Constrained by 15 GB RAM on a 1.65 x 10^9-entry sparse matrix, the Executor developed a grep-and-awk streaming pipeline processing 28 DA marker genes in 138 seconds under 1 GB RAM. Subject-level pseudobulk expression across 22 subjects (4 control, 18 LBD) identified three gene classes: control-enriched markers lost in LBD (CALB1 FC = 0.04; SOX6 FC = 0.13; DRD2 FC = 0.09), stable markers (NR4A2 FC = 0.75), and zero-inflation artifacts (SLC6A3 FC = 109.5; EN1 FC = 105.2—identified by the Critic agent). CALB1 depletion in PD SNc corresponds to 36-year prior art (Yamada et al., 1990), providing statistically independent subject-level validation. Critically, the Literature agent corrected an autonomous "dorsal-tier vulnerability" formulation to "ventral-tier vulnerability with dorsal-tier relative sparing"—a directional error introduced by the absence of literature context during autonomous analysis. In one notable incident, the Critic itself committed a methodological error (two-tailed vs. one-tailed MWU), caught and retracted by the Reporter agent.

5. Discussion

5.1 Complementary Architecture

The three layers form a complementary capability stack where each layer builds on the reliability established by the layer below. Layer 1's near-perfect padj correlation and >90% pathway overlap provide the reliability baseline that justifies trust in agent-driven analysis. Layer 2's multi-backend scoring architecture and context annotation demonstrate how agents can transcend simple command execution to become domain-specific design tools—the explicit layering of CRISPR Agent's pipeline directly inspired SciOS's modular audit architecture. Layer 3's FORUM gate chain and multi-agent audit network realize the ultimate capability: autonomous generation of novel, validated scientific knowledge.

A notable emergent property is cross-layer error correction. In Layer 1, agent-pipeline discrepancies exposed data format and filtering issues, motivating explicit computational boundaries in Layer 2. In Layer 2, cross-backend scoring comparisons revealed the limitations of individual scoring methods, motivating multi-modal validation in Layer 3. In Layer 3, the multi-agent audit network caught errors even in the auditor itself—demonstrating that no single layer of quality control is sufficient. This "meta-audit" capability—checking the checker—is among the most distinctive emergent properties of the three-layer framework.

5.2 Limitations and Future Directions

(1) Cross-layer communication. The three layers currently operate in independent projects without automated cross-layer information transfer. Future versions should allow Layer 1 reliability metrics to automatically calibrate confidence thresholds in Layer 3.

(2) Error propagation detection. The current framework lacks systematic mechanisms for detecting error propagation from lower-layer agents to higher-layer findings. While multi-layer audit partially mitigates this, structured error-propagation tracing is needed.

(3) Human intervention gradient. The three layers require different magnitudes of human intervention—from minimal prompting in Layer 1 to user-defined scientific questions in Layer 3. Engineering a smooth autonomy gradient remains a key challenge.

(4) Domain expansion. The framework has been validated primarily on genomics and transcriptomics tasks. Extension to proteomics, metabolomics, and imaging data is an important next step for generality assessment.

5.3 Conclusion

This work demonstrates a complete evolution from pipeline automation to scientific discovery through a three-layer agent-driven bioinformatics framework. The layers are not isolated experiments—they represent a continuous capability spectrum of the same technology stack. Foundation validation provides trust, domain tools provide design patterns, and the scientific discovery system realizes the ultimate goal: machines that do not merely run analyses, but ask questions, audit findings, and discover new knowledge. The three-layer architecture collectively establishes methodological foundations, design principles, and empirical evidence for constructing next-generation autonomous bioinformatics systems.

Acknowledgments

Layer 1 used TCGA LIHC data from GDC and E. coli sequencing data (SRR1770413) from NCBI SRA. Layer 2 CRISPR Agent scoring validation used public GUIDE-seq, CIRCLE-seq, and CHANGE-seq datasets. Layer 3 SciOS used publicly available single-cell and spatial transcriptomic data from the Allen Institute for Brain Science (GSE115746), Broad Institute (GSE178265), and Lieber Institute for Brain Development (GSE307990). All computation performed under 15 GB RAM Linux without GPU acceleration. The memory constraints forced methodological innovations (grep-and-awk streaming pipeline, genotype-agnostic pivot strategy) that ultimately strengthened analytical rigor across multiple projects.

References

[1] Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. Science 339, 819-823 (2013).

[2] Jinek, M. et al. A Programmable Dual-RNA-Guided DNA Endonuclease. Science 337, 816-821 (2012).

[3] Hsu, P. D. et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. 31, 827-832 (2013).

[4] Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage. Nat. Biotechnol. 33, 187-197 (2015).

[5] Bae, S., Park, J. & Kim, J.-S. Cas-OFFinder. Bioinformatics 30, 1473-1475 (2014).

[6] Doench, J. G. et al. Optimized sgRNA design to maximize activity. Nat. Biotechnol. 34, 184-191 (2016).

[7] Lorenz, R. et al. ViennaRNA Package 2.0. Algorithms Mol. Biol. 6, 26 (2011).

[8] Quinlan, A. R. & Hall, I. M. BEDTools. Bioinformatics 26, 841-842 (2010).

[9] Cameron, P. et al. Mapping the genomic landscape of CRISPR-Cas9 cleavage. Nat. Methods 14, 600-606 (2017).

[10] Lazzarotto, C. R. et al. CHANGE-seq. Nat. Biotechnol. 38, 1317-1327 (2020).

[11] Tsai, S. Q. et al. CIRCLE-seq. Nat. Methods 14, 607-614 (2017).

[12] Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. Nature 563, 72-78 (2018).

[13] Hodge, R. D. et al. Conserved cell types with divergent features. Nature 573, 61-68 (2019).

[14] Kamath, T. et al. Single-cell genomic profiling of human dopamine neurons. Nat. Neurosci. 25, 588-595 (2022).

[15] Huuki-Myers, L. A. et al. APOE E4 risk converges on an oligodendrocyte subtype. bioRxiv (2025).

[16] Pandey, A. C. et al. A CRISPR/Cas9-based enhancement of single-cell transcriptomics. Nat. Commun. 16, 4664 (2025).

[17] Yamada, T. et al. Relative sparing of calbindin-D28K neurons in PD. Brain Res. 526, 303-307 (1990).

[18] Damier, P. et al. Nigrosomes and the nigral matrix. Brain 122, 1421-1436 (1999).

[19] Yamazaki, Y. et al. APOE and Alzheimer disease. Nat. Rev. Neurol. 15, 501-518 (2019).

[20] Bakken, T. E. et al. Comparative cellular analysis of motor cortex. Nature 598, 111-119 (2021).

[21] Pereira Luppi, M. et al. Sox6 distinguishes DA neuron populations. Cell Rep. 37, 109975 (2021).

[22] Dykxhoorn, D. M. et al. Ancestral genomic differences in oligodendroglia. Res. Sq. (2024).