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.

Human-Governed Multi-Agent AI for Biological Hypothesis Generation: An Evidence-Grounded Framework Using JUMP Cell Painting Profiles

Stephen Kamran

kamran.msbf@gmail.com

Government College University, Lahore https://orcid.org/0009-0000-3063-5611

Farah Saman

University of Exeter https://orcid.org/0000-0003-0933-4368

Research Article

Keywords: Agentic AI, autonomous scientific discovery, Cell Painting, CRISPR, evidence-grounded AI, JUMP, multi-agent systems, morphological profiling, retrieval-augmented generation, human-in-the-loop

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11555250/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.

Human-Governed Multi-Agent AI for Biological Hypothesis Generation: An Evidence-Grounded Framework Using JUMP Cell Painting Profiles

Stephen Kamran
Government College University Lahore
Lahore, Pakistan
kamran.msbf@gmail.com
ORCID: 0009-0000-3063-5611

Farah Saman
University of Exeter
Exeter, United Kingdom
fs531@exeter.ac.uk
ORCID: 0000-0003-0933-4368

Abstract— High-dimensional cellular phenotypes can motivate biological hypotheses, but similarity alone cannot establish mechanism or causal validity. DISCOVERY-AGENT links exploratory JUMP Cell Painting evidence with evidence-grounded hypothesis generation, independent critique and bounded human-review routing. An initial 65-record CRISPR/ORF analysis across nine gene–modality tests found four FDR-significant CRISPR plate-level associations (AKT1, BRCA1, MTOR and TP53), none among evaluated ORF genes. A distinct later cross-source compound-identity benchmark used 614 morphological features and achieved AUROC 0.6142 on 1,250 held-out compounds with training-only preprocessing; this is identity discrimination, not mechanism-of-action classification. Three structured hypotheses underwent LLM critique and revision, while eleven gene-pair evidence assessments yielded no FDR-significant mechanistic finding (minimum q = 0.2037). Archived governance checks passed 1,021/1,021 predefined tests; nevertheless, the initial live-agent run contained two execution errors among five cases, with both cases responding on fresh retest. These findings demonstrate feasibility of auditable evidence processing and modest cross-source morphology signal, not validated autonomous discovery, prospective safety, or completed expert review. The work delineates practical requirements for independent biological validation and reproducible scientific agents.

Index Terms— *Agentic AI, autonomous scientific discovery, Cell Painting, CRISPR, evidence-grounded AI, JUMP, multi-agent systems, morphological profiling, retrieval-augmented generation, human-in-the-loop.*

I. INTRODUCTION

The convergence of high-content cellular imaging, genetic perturbation screening, and artificial intelligence (AI) is reshaping biological discovery by enabling systematic investigation of cellular responses, gene functions, and therapeutic mechanisms. Modern experimental platforms generate multidimensional phenotypic measurements across chemical and genetic interventions, creating opportunities for computationally assisted scientific discovery. However, translating complex morphological observations into reproducible, mechanistically meaningful, and experimentally testable hypotheses remains a fundamental challenge. Predictive performance and phenotypic similarity cannot independently establish molecular causality, biological validity, or therapeutic relevance [1], [2]–[4].

Cell Painting has emerged as a powerful approach for high-dimensional cellular phenotyping through multiplexed fluorescence microscopy. The Joint Undertaking for Morphological Profiling (JUMP) provides extensive chemical and genetic perturbation resources, enabling systematic characterization of cellular responses and functional relationships [5]–[7]. Recent advances in morphological mapping, self-supervised representation learning, confounder-aware modeling, and enhanced imaging protocols have strengthened phenotype characterization [8]–[12]. Complementary investigations demonstrate the potential of Cell Painting for drug-response profiling, mechanism-of-action inference, and integration with transcriptomic, chemical, and proteomic information [13]–[21]. Nevertheless, experimental heterogeneity, plate effects, technical confounding, and biological variability complicate interpretation. Consequently, morphological associations require rigorous statistical evaluation before supporting mechanistic hypotheses.

Cellular foundation models and perturbation-response prediction have further expanded computational capabilities. Recent approaches employ multimodal alignment, generative modeling, and learned cellular representations to characterize intervention-associated biological changes [1], [22], [23], [24], [25], [26], [27], [28]. However, evidence that sophisticated deep-learning models do not consistently outperform simpler baselines raises concerns regarding generalization, evaluation rigor, and biological interpretability [2], [3]. Trustworthy virtual-cell research therefore requires reproducible experimentation, uncertainty assessment, and independent validation rather than reliance on predictive accuracy alone [4]. These limitations highlight the need for computational frameworks that connect statistical evidence with biologically grounded reasoning.

Simultaneously, autonomous scientific-discovery systems are emerging as a promising paradigm for coordinating complex research workflows. Multi-agent architectures distribute analytical responsibilities across specialized components responsible for evidence interpretation, literature retrieval, hypothesis formulation, critical reasoning, and experimental planning [29]–[34]. Developments in agentic bioinformatics, biological intelligence, and autonomous research ecosystems further illustrate the potential of coordinated AI-assisted scientific investigation [35]–[40]. Biomedical knowledge graphs and retrieval-augmented generation (RAG) offer complementary mechanisms for integrating documented biological relationships with traceable scientific evidence [41]–[46]. Nevertheless, generative agents may produce plausible but unsupported explanations, propagate retrieval errors, or reinforce incorrect conclusions through artificial consensus. Scientific reliability therefore requires evidence provenance, independent verification, calibrated uncertainty, and human oversight.

Despite these developments, a critical methodological gap persists between high-content cellular profiling and autonomous biological discovery. Existing research addresses morphological representation learning, perturbation prediction, drug-target identification, knowledge-grounded reasoning, batch correction, and uncertainty-aware validation [19]–[21], [28], [47]–[53]. However, these capabilities have not established a sufficiently validated, integrated workflow connecting reproducible JUMP-derived phenotypic

evidence with independently testable biological hypotheses. Specifically, the distinction between statistically supported morphological associations, computationally inferred explanations, and experimentally confirmed mechanisms remains insufficiently operationalized in autonomous discovery pipelines. Without explicit evidentiary boundaries, computational plausibility risks being misinterpreted as scientific confirmation.

To address this challenge, this study introduces **DISCOVERY-AGENT**, an evidence-grounded framework for autonomous scientific reasoning using JUMP-derived cellular perturbation data. The investigation comprises two distinct components. First, a reproducible exploratory analysis evaluates CRISPR and open reading frame (ORF) perturbations involving AKT1, BRCA1, EGFR, MTOR, and TP53 through plate-level statistical inference and false discovery rate correction. Four of five evaluated CRISPR gene tests exhibit significant morphological associations at an adjusted threshold of 0.05, whereas EGFR does not; none of the evaluated ORF gene tests reaches significance. These findings provide directly auditable phenotypic evidence but do not establish molecular causality, drug-target relationships, or autonomous discovery performance.

Second, DISCOVERY-AGENT specifies a modular architecture integrating statistical evidence extraction, biological knowledge retrieval, specialized hypothesis-generation agents, independent evidence critique, uncertainty assessment, and human-in-the-loop verification. The proposed system prioritizes evidence traceability, competing biological explanations, and explicit validation requirements. Its governing principle is that autonomous scientific reasoning must remain accountable to reproducible observations and independently verifiable knowledge. Accordingly, generated hypotheses are treated as provisional scientific propositions rather than established biological discoveries. The architecture remains a testable research proposal requiring dedicated experimental benchmarking.

The study makes four principal contributions: **(1)** a reproducible exploratory assessment of selected JUMP-derived genetic perturbations using plate-level inference and multiple-testing correction; **(2)** an explicit evidentiary distinction between observed phenotypic associations and unverified mechanistic interpretations; **(3)** a proposed multi-agent architecture integrating biological RAG, hypothesis generation, evidence verification, uncertainty awareness, and bounded human oversight; and **(4)** a future evaluation framework emphasizing biological validity, calibration, evidence traceability, reproducibility, and comparison against appropriate computational baselines.

By integrating auditable cellular phenotyping with verification-centered agentic reasoning, this research establishes a methodological foundation for trustworthy AI-assisted biological hypothesis generation. Rather than replacing experimental biology, DISCOVERY-AGENT aims to support a transparent progression from cellular observations to scientifically testable explanations. The following sections present related research, experimental methodology, the proposed architecture, empirical findings, and future validation requirements.

II. RELATED WORK

A. Cell Painting and perturbation phenotyping

High-content cellular imaging has become an important approach for investigating biological responses to genetic and chemical interventions. The JUMP Cell Painting resources provide large-scale morphological measurements that support comparative phenotyping, functional characterization, and perturbation-associated discovery [5]–[7]. Chandrasekaran et al. [8] extended morphological mapping to genetic underexpression and overexpression, strengthening the potential for identifying functional relationships among perturbations. However, high-dimensional phenotypic similarity remains an indirect indicator of biological relatedness and cannot independently establish shared molecular mechanisms.

Recent methodological advances have focused on improving representation quality and controlling experimental confounding. Yan et al. [9] investigated contrastive and domain-adversarial learning for correcting technical effects, while Kim et al. [10] demonstrated the value of self-supervised image representations for morphological profiling. Papanastasiou et al. [11] further addressed confounding within foundation-model-based phenotype analysis. Complementary developments in expanded Cell Painting protocols [12] and context-sensitive phenotyping [13], [14] highlight the importance of measurement diversity and biological context. Collectively, these studies suggest that representation quality must be evaluated alongside robustness to experimental variation.

Cell Painting has also demonstrated applications in compound prioritization, gene-function inference, drug screening, and mechanism-of-action investigation [15]–[18]. Multimodal approaches further connect morphological measurements with transcriptomic, chemical, and proteomic information [19]–[21]. Such integration is relevant to biological hypothesis generation because independent molecular evidence can help distinguish phenotypic convergence from mechanistic correspondence. Nevertheless, cross-modal agreement remains suggestive rather than conclusive without appropriate experimental confirmation.

Cellular foundation models have expanded perturbation-response modeling through learned representations, multimodal alignment, and generative prediction [1], [22], [23], [24], [25], [26], [27], [28]. However, Ahlmann-Eltze et al. [2] reported that sophisticated deep-learning approaches do not necessarily outperform simple linear baselines in gene perturbation prediction. Broader methodological discussions similarly emphasize experimental context, generalization, and evaluation rigor [3], [4]. These findings challenge the assumption that increasing model complexity necessarily improves scientific validity. For DISCOVERY-AGENT, they motivate an evidence-first approach in which morphological associations are established statistically before being incorporated into mechanistic reasoning.

B. Agentic discovery and evidence grounding

Autonomous scientific-discovery systems increasingly employ coordinated AI components to support complex research processes. Ghareeb *et al.* [29] investigated multi-agent scientific discovery, while recent biological-agent research has examined specialized computational roles across biomedical investigation [30]–[32]. Related developments in agentic bioinformatics and autonomous research workflows demonstrate growing interest in systems capable of integrating analytical execution, information retrieval, hypothesis formulation, and iterative reasoning [33]–[40].

A central advantage of multi-agent architectures is functional specialization. Distinct agents can be assigned responsibility for statistical interpretation, biological context retrieval, candidate hypothesis generation, critical review, and experimental

prioritization. However, specialization does not guarantee independent reasoning or scientific correctness. Agents sharing the same model, retrieval corpus, or assumptions may reproduce correlated errors, creating misleading consensus. Consequently, agreement among agents should be interpreted as an internal computational signal rather than independent biological validation.

Knowledge-grounded reasoning offers a potential mechanism for addressing these limitations. Yang *et al.* [41] examined traceable biomedical hypothesis generation through causal reasoning, while Borrego *et al.* [42] investigated hypothesis generation using scientific knowledge graphs. Biomedical multi-agent retrieval [43], GraphRAG [44], federated knowledge retrieval [45], and adaptive knowledge-graph-guided verification [46] provide complementary approaches for connecting generated explanations with external evidence. These developments are particularly relevant to JUMP, where morphological observations may require contextualization through gene functions, protein interactions, molecular pathways, and compound-target relationships.

Nevertheless, retrieval-grounded generation introduces its own limitations. Retrieved documents may be incomplete, outdated, contradictory, or insufficiently specific to the experimental context. Knowledge-graph relationships may represent associations rather than causal mechanisms, while plausible generated explanations can remain unsupported by the underlying perturbation data. DISCOVERY-AGENT therefore adopts evidence traceability, competing-hypothesis assessment, explicit uncertainty, and human review as proposed architectural requirements. Existing agentic studies motivate these design choices but do not establish their effectiveness on the present JUMP exploratory dataset.

C. Robust evaluation

Reliable perturbation modeling requires evaluation procedures that distinguish genuine biological signal from technical artifacts, statistical instability, and information leakage. Batch-effect correction remains important in high-content screening [9], [11], [47], while comparative analyses of cellular representations emphasize that apparent predictive improvements must be interpreted relative to appropriate baselines [1], [26], [27], [2]. Evaluation must also consider whether train–test partitions preserve experimental independence and whether observed performance generalizes across plates, perturbations, and biological contexts.

Drug-target identification and mechanism-of-action inference introduce additional validation requirements. Network-based approaches [48], AI-assisted target assessment [49], interpretable pharmacodynamic knowledge graphs [50], and graph-based mechanistic discovery [51] demonstrate the potential of combining computational predictions with biological relationships. However, a predicted target or pathway remains a hypothesis until supported by independent evidence.

Uncertainty-aware decision systems provide another important methodological foundation. Conformal validation and human-in-the-loop deferral [52] illustrate approaches for identifying predictions requiring additional review. Retrieval-based evaluation of Cell Painting representations [53] further highlights the importance of biologically meaningful similarity rather than superficial embedding proximity.

Research synthesis: Existing studies advance phenotype representation, perturbation prediction, biological knowledge retrieval, and autonomous reasoning, but they do not collectively validate the proposed DISCOVERY-AGENT workflow. The present study therefore separates reproducible exploratory phenotypic inference from a proposed, verification-centered multi-agent architecture. Future benchmarking must establish whether that architecture improves hypothesis validity, evidence traceability, uncertainty calibration, and experimental prioritization relative to simpler alternatives.

III. MATERIALS AND METHODS

A. Source artifacts and data provenance

This study adopts a reproducibility-oriented computational research design combining exploratory phenotypic analysis with a proposed evidence-grounded multi-agent discovery architecture. The biological foundation is the Joint Undertaking for Morphological Profiling (JUMP), a large-scale Cell Painting initiative providing morphological measurements associated with chemical and genetic perturbations [5]–[7]. The scientific relevance of these resources has been further established through systematic morphological mapping of genetic interventions [8] and advances in Cell Painting-based functional characterization [16].

The underlying JUMP ecosystem provides publicly accessible dataset descriptions, perturbation metadata, imaging resources, and associated computational workflows through the JUMP Consortium repository and the Cell Painting Gallery. The foundational JUMP publication is available at https://doi.org/10.1038/s41592-024-02241-6. These resources establish the provenance of the broader experimental platform and provide access routes for future independent replication.

The present empirical investigation uses a limited set of previously generated JUMP-derived perturbation summaries. The archived analytical outputs comprise CRISPR and open reading frame (ORF) phenotype-result tables, plate-level statistical validation results, and modality-specific visualizations. The recorded analysis configuration specifies a random seed of 42 and identifies the statistical procedure as full-feature, plate-level exploratory sign testing. Independent biological replication has not been established.

Importantly, the available analytical outputs do not constitute the complete original JUMP morphological-profile matrices. Therefore, the present study reports the archived statistics without claiming to have independently reconstructed every upstream feature transformation, normalization procedure, or null-score calculation from raw cellular measurements.

B. Pilot sample and outcomes

The exploratory analysis considers two genetic perturbation modalities: clustered regularly interspaced short palindromic repeats (CRISPR)-associated interventions and ORF-based gene overexpression. These modalities are scientifically relevant because perturbation direction and experimental implementation may produce distinct cellular responses, even when targeting related biological processes [8], [16].

The CRISPR analysis contains 30 summarized records distributed across five genes: AKT1, BRCA1, EGFR, MTOR, and TP53. The ORF analysis contains 35 records involving AKT1, BRCA1, EGFR, and TP53. The recorded feature dimensionality differs by modality, with 259 morphological features represented in the CRISPR analytical records and 722 in the ORF records.

Each record identifies the perturbation modality, target gene, experimental plate, phenotype score, excess score, and feature count. These records are derived analytical summaries rather than individual-cell observations. Consequently, the number of records must not be interpreted as the number of independent biological replicates.

The inclusion of CRISPR and ORF perturbations provides an exploratory basis for examining modality-dependent morphological responses. However, the sample is not designed to establish generalizable differences between gene knockout and overexpression mechanisms. Recent research on genetic perturbation maps and multimodal cellular modeling illustrates the importance of considering perturbation context and measurement representation when interpreting such comparisons [8], [19], [23], [28].

C. Plate-level testing and multiplicity

The empirical objective is to identify perturbation-associated morphological evidence while minimizing overinterpretation of technical or statistically unstable effects. Contemporary morphological profiling research emphasizes representation quality, confounding control, and sensitivity to batch-specific variation [9]–[11], [47]. Accordingly, the analysis evaluates summarized phenotype effects at the experimental-plate level rather than treating repeated measurement rows as independent inferential units.

For each evaluated gene–modality combination, the archived statistical output reports the number of profiles, number of represented plates, number of plates with positive effects, mean absolute standardized phenotype score, median excess score, and exact sign-test probability.

The exact sign test evaluates whether the observed direction of the plate-level effects is inconsistent with a null hypothesis of equally likely positive and negative signs, subject to the test's assumptions. In particular, inferential validity depends on the independence and appropriate construction of the plate-level observations. Because the exact construction of the underlying plate-level excess scores is not fully recoverable from the available summary tables, these equations describe the statistical framework rather than an independently reimplemented analytical pipeline.

To address multiple hypothesis testing, the reported results employ the Benjamini–Hochberg false discovery rate adjustment. Statistical significance is assessed using the archived adjusted values, with ($q<0.05$) as the reported discovery criterion.

Particular care is required when interpreting ORF observations because multiple summarized records can originate from the same plate. Treating these records as independent replicates would risk pseudoreplication and potentially misleading statistical certainty.

The analysis therefore distinguishes statistical evidence of a morphological association from biological replication, causal inference, and confirmed molecular mechanism. This conservative interpretation is consistent with recent recommendations for perturbation-model evaluation and trustworthy virtual-cell development [26], [2], [4], [53].

D. Proposed DISCOVERY-AGENT workflow

Building upon the exploratory phenotypic analysis, DISCOVERY-AGENT is specified as an evidence-grounded, multi-agent scientific reasoning framework. The architecture draws methodological inspiration from autonomous research systems, biomedical AI agents, and specialized scientific workflows [29]–[40]. Its evidence-integration strategy is further informed by traceable hypothesis generation, biomedical knowledge graphs, and retrieval-augmented generation (RAG) [41]–[46].

The proposed workflow contains six interconnected stages.

Stage 1 — Provenance-aware data ingestion: Perturbation profiles, experimental metadata, plate identifiers, controls, and analysis provenance are organized into auditable records. This stage establishes the relationship between individual computational outputs and their underlying experimental conditions.

Stage 2 — Plate- and batch-aware phenotypic analysis: Morphological evidence is characterized using appropriate normalization, aggregation, effect estimation, and statistical validation procedures. Confounder-aware representation learning and batch-correction methods provide relevant methodological precedents [9]–[11], [47]. Only analytically characterized evidence proceeds to downstream reasoning.

Stage 3 — Biological knowledge retrieval: Gene identifiers, perturbation signatures, and candidate pathway relationships are used to formulate queries against documented biological evidence. A future implementation may combine scientific literature retrieval, curated molecular databases, and knowledge-graph representations. This stage is motivated by biomedical RAG, structured hypothesis-generation approaches, and graph-based reasoning [41]–[46], [48], [51].

Stage 4 — Multi-agent hypothesis generation and verification: A hypothesis-generation agent proposes candidate biological explanations linking observed phenotypes to plausible molecular processes. A separate verification agent examines whether retrieved evidence supports the proposed interpretation, identifies contradictory findings, and assesses alternative explanations. This separation is intended to discourage unsupported mechanistic assertions and draws on recent work in traceable scientific reasoning and autonomous discovery [29], [31], [41], [42].

Stage 5 — Uncertainty-aware prioritization: Candidate hypotheses are evaluated according to evidence completeness, statistical support, contradictions, and unresolved assumptions. Hypotheses with weak or inconsistent support are deferred rather than automatically accepted. Uncertainty-aware validation and selective human review provide relevant methodological precedents [26], [4], [52].

Stage 6 — Human-in-the-loop scientific review: A domain expert reviews the surviving hypotheses, supporting evidence, competing interpretations, and proposed experimental tests. The final disposition is recorded as approval for follow-up, revision, deferral, or rejection. Approval at this stage indicates permission for further investigation, not confirmation of a biological mechanism.

The architecture is designed to preserve a structured evidence trail containing the originating perturbation, statistical result, source-linked biological evidence, counterevidence, uncertainty rationale, agent-review outcome, and human decision. These specifications constitute proposed system requirements and have not been validated as operational performance outcomes in the present exploratory investigation. The proposed information flow is illustrated in Fig. 1, and its stage-level inputs and verification gates are summarized in Table I.

TABLE I. Proposed DISCOVERY-AGENT functional architecture (not experimentally validated).

Stage	Input	Output / verification gate
1. Data provenance	JUMP metadata, profiles, controls	Traceable experimental records
2. Phenotypic evidence	Perturbations and plate IDs	Plate-aware effects and statistical evidence
3. Biological retrieval	Gene and pathway queries	Source-linked evidence bundle
4. Hypothesis generation	Phenotype effects and evidence	Candidate, provisional hypotheses
5. Verification and triage	Candidates and counterevidence	Supported, contradicted, or uncertain
6. Human review	Audited candidate and rationale	Test, revise, defer, or reject

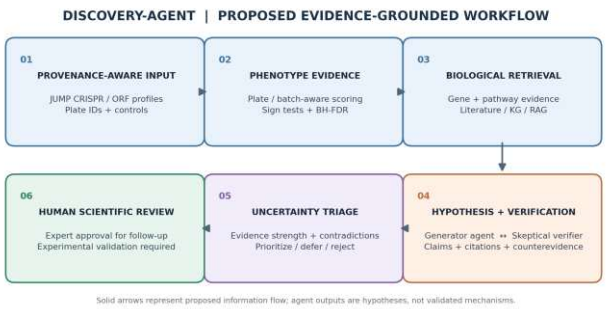


Fig. 1. Proposed six-stage DISCOVERY-AGENT architecture showing provenance-aware JUMP ingestion, plate-level analysis, biological retrieval, paired hypothesis generation and verification, uncertainty triage, and human review (conceptual design; not experimentally validated).

E. Reproducibility and Evaluation Boundaries

The study maintains a formal separation between completed statistical analysis and proposed autonomous scientific reasoning. The empirical component provides archived gene-level summaries, plate-level inference results, and visualizations. The architectural component specifies a candidate workflow for future implementation and evaluation.

Reproducibility is tiered by experiment. The 65-record plate-level pilot is inspectable from archived gene-level results, test statistics and plots, but it cannot be fully recomputed from raw images without the exact profile extracts, plate/control mappings, upstream normalization and null-score construction. The later 614-feature compound-identity experiment has an archived Kaggle analysis notebook, intermediate Parquet representations, split definitions, aggregate metrics and result files; independent re-execution additionally requires the versioned source data and executable preprocessing dependencies. External LLM calls require provider access, model-version records, prompts, inference settings and quota availability. Reproducing saved result tables is not equivalent to a successful clean-environment end-to-end run.

The cross-source test used source-specific fitting and a compound-disjoint held-out set; this limits identity leakage but does not demonstrate inter-laboratory biological generalization or mechanism-of-action prediction. Agent workflow evidence includes three revision-traceable provisional hypotheses, eleven gene-pair statistical assessments, rule-based governance checks and five live execution cases. Without independent mechanistic truth labels or paired, controlled retrieval-only, single-agent and multi-agent runs, we cannot estimate autonomous hypothesis accuracy or comparative agent superiority. Reproducibility and future validation must report these evaluation layers separately.

Accordingly, the present research establishes a reproducible reporting basis for a limited exploratory analysis and defines the verification requirements for subsequent experimental assessment of autonomous biological hypothesis generation.

IV. RESULTS

A. Exploratory plate-level phenotype signals

Among five evaluated CRISPR gene tests, four achieved Benjamini–Hochberg-adjusted significance ($q < 0.05$): AKT1 ($q = 0.039062$), BRCA1 ($q = 0.004883$), MTOR ($q = 0.039062$), and TP53 ($q = 0.039062$). BRCA1 showed the strongest evidence, with positive excess scores on all 10 represented plates ($p = 0.000977$). AKT1, MTOR, and TP53 each showed positive effects on all five represented plates ($p = 0.031250$). EGFR did not meet the adjusted criterion (3/5 positive plates; $q = 0.500000$). These findings represent exploratory plate-level associations rather than confirmed gene function or causal mechanisms. Table II reports the complete results, while Figs. 2 and 3 display modality-specific mean absolute phenotype-score summaries.

The ORF modality demonstrated a different pattern. None of the four evaluated ORF gene tests satisfied the multiple-testing criterion. AKT1 and EGFR each exhibited positive effects in eight of ten plates, but their adjusted significance values remained above the prespecified threshold ($q = 0.072917$). BRCA1 also failed to achieve significance ($q = 0.072917$), whereas TP53 showed comparatively weaker evidence ($q = 0.500000$). These observations indicate that the detectable phenotypic evidence differed between the evaluated modalities and genes, although the limited sample does not permit generalizable conclusions about differences in CRISPR- and ORF-induced biological responses. The two modality-specific mean absolute deviation summaries are shown in Figs. 2 and 3, respectively; bar height should not be interpreted as statistical significance.

TABLE II. Final plate-level validation, transcribed from the supplied CSV (rounded for display)

Perturbation	Positive	p	FDR q
CRISPR / AKT1	5/5	0.0312	0.0391
CRISPR / BRCA1	10/10	0.0010	0.0049
CRISPR / EGFR	3/5	0.5000	0.5000
CRISPR / MTOR	5/5	0.0312	0.0391
CRISPR / TP53	5/5	0.0312	0.0391
ORF / AKT1	8/10	0.0547	0.0729
ORF / BRCA1	5/5	0.0312	0.0729
ORF / EGFR	8/10	0.0547	0.0729
ORF / TP53	3/5	0.5000	0.5000

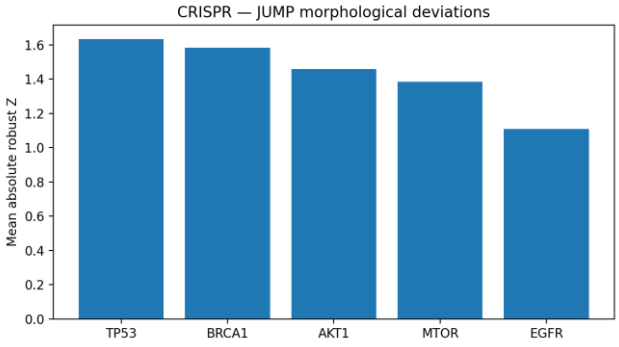


Fig. 2. Mean absolute morphological deviation across selected CRISPR gene perturbations (exploratory analysis).

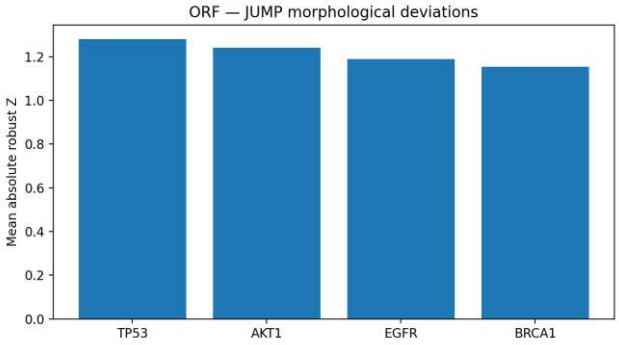


Fig. 3. Mean absolute morphological deviation across selected ORF gene perturbations (exploratory analysis).

The observed CRISPR and ORF phenotype patterns are broadly consistent with findings from Chandrasekaran et al. [8], who investigated morphological responses to genetic underexpression and overexpression in the JUMP Cell Painting resource. Their genome-scale investigation reported detectable phenotypic activity for approximately 70% of evaluated CRISPR-targeted genes and 56% of evaluated ORF-overexpressed genes. The study also demonstrated that genetic perturbations can yield biologically informative morphological signatures, while emphasizing the importance of rigorous preprocessing and technical-artifact control.

An additional methodological correspondence is evident in the processed feature dimensions: Chandrasekaran et al. [8] reported 259 features for CRISPR profiles and 722 for ORF profiles, matching the dimensions represented in the present analytical records. This agreement supports the interpretation that the exploratory analysis uses feature representations consistent with the published JUMP morphological-profiling framework; however, feature-count agreement alone does not verify the exact upstream source files or preprocessing history.

Importantly, the present study's four significant CRISPR tests out of five and zero significant ORF tests out of four should not be interpreted as direct replications of the published 70% and 56% phenotypic-activity rates. The studies differ substantially in sample scope, statistical procedures, and criteria for detecting phenotypic activity. Their agreement is therefore methodological and qualitative rather than an independent confirmation of gene-specific outcomes.

Related research further supports the biological utility of morphological perturbation measurements. Sivanandan et al. [16] investigated self-supervised analysis of pooled CRISPR Cell Painting screens for gene-function inference, illustrating the potential of image-derived perturbation signatures for biological characterization. Confounder-aware representation learning [9], [11] and retrieval-based assessment of morphological similarity [53] likewise emphasize that biologically meaningful phenotype detection depends on robust feature processing and careful interpretation.

Collectively, these comparisons position the present findings within established Cell Painting research while preserving the distinction between a limited exploratory statistical signal and independently validated biological mechanism.

B. What the artifacts do not establish

Although the exploratory analysis provides evidence of selected perturbation-associated morphological signals, the results do not establish the performance of the proposed DISCOVERY-AGENT architecture. The archived development notebook contains limited feasibility traces but no controlled head-to-head comparison between autonomous agents and conventional computational baselines, validated compound–CRISPR matching outcomes, or experimentally confirmed agent-generated mechanistic hypotheses.

Accordingly, agent-level performance measures—including hypothesis-generation accuracy, precision, recall, F1-score, area under the receiver operating characteristic curve (AUROC), calibration, and scientific discovery success rate—cannot be estimated reliably from the present results. No statistically supported superiority claims are therefore made for DISCOVERY-AGENT over single-agent systems, retrieval-only approaches, foundation models, or alternative large language model providers.

This distinction is particularly important given recent concerns about perturbation-model benchmarking. Ahlmann-Eltze et al. [2] demonstrated that increasingly complex deep-learning approaches do not necessarily outperform simpler predictive baselines, reinforcing the need for controlled comparative evaluation. Related work on perturbation benchmarking [1], [26], [27] and uncertainty-aware human-in-the-loop validation [52] further motivates independent test sets, evidence-grounding assessments, leakage control, and explicit uncertainty reporting.

Exploratory notebook executions involving external AI services are consequently treated as development-stage computational experiments. Without frozen prompts, model identifiers, inference settings, responses, evaluation labels, and reproducible execution records, these observations cannot support formal statistical comparisons between agent configurations or language model providers.

Taken together, the archived plate-level analyses provide exploratory morphological associations, while the separate cross-source experiment quantifies modest compound-identity discrimination. The agent and governance traces document executable evidence processing and constrained decisions, not independent biological validation or proven advantages over simpler systems.

C. Exploratory Agent Feasibility (Development-Stage Audit)

The archived development notebook also records a limited execution audit for five gene-pair cases: BRCA1–TP53, AKT1–MTOR, and AKT1–BRCA1 (CRISPR); and AKT1–TP53 and AKT1–EGFR (ORF). These runs examine the feasibility of structured AI responses, biomedical literature retrieval, and evidence triage. They are distinct from the plate-level phenotype-

significance tests in Table II and do not constitute a controlled evaluation of the complete proposed DISCOVERY-AGENT architecture.

Across the five recorded cases, the initial Groq requests were marked successful and all five parsed outputs were valid JSON. The initial Gemini execution succeeded for three of five cases; two calls returned HTTP 503 errors, including on recorded retry attempts. These provider-specific failures did not terminate the exploratory workflow. The archived notebook subsequently records a successful Together AI API connection (HTTP 200), a successful inference test using meta-llama/Llama-3.3-70B-Instruct-Turbo (HTTP 200), and two subsequent case-level verification calls for AKT1–TP53 and AKT1–EGFR (both HTTP 200; both verdicts "insufficient"). The case-level audit ultimately records five completed verification statuses, with human review required in all five cases; the two later cases are explicitly marked SUCCESS_TOGETHER. PubMed searches succeeded for all five gene-pair cases, and the final quality-control audit marked 25 of 25 article-level evidence assessments "INSUFFICIENT" for their specific hypotheses. An earlier case-level "mixed" verdict for AKT1–MTOR concerns a different level of evaluation and is not inconsistent with the later article-level audit.

TABLE III. Archived exploratory agent execution and evidence-review audit (five cases; not a comparative benchmark).

Observed item	Archived result	Interpretation
Gene-pair cases	5 (3 CRISPR; 2 ORF)	Small exploratory execution set
Groq execution / valid JSON	5/5; 5/5	Request completion and output syntax only
Initial Gemini execution	3/5 successful; 2 HTTP 503	Initial provider attempts only; not final case success
PubMed case retrieval	5/5 successful	Five PMIDs retained per case
Article-level review after QC	25/25 insufficient	No sufficient direct support established in this audit
Human review required	5/5 cases	Escalation flags, not validated triage performance
Together AI continuation	API/model checks HTTP 200; 2/2 follow-up case calls HTTP 200	Llama-3.3-70B-Instruct-Turbo; two "insufficient" case verdicts

The archived notebook does not supply independent mechanistic ground-truth labels, a frozen test set, or a paired retrieval-only/single-agent/multi-agent comparison. Accordingly, JSON validity, provider availability, retrieval counts, and evidence-review verdicts are reported as operational observations rather than hypothesis accuracy, evidence-grounding precision, sensitivity, or scientific-discovery success. The present audit supports feasibility of certain development components, not validated agent effectiveness or biological confirmation. Successful Together AI execution does not measure ground-truth biological accuracy or establish superiority over another model provider.

D. Additional JUMP profile benchmark and governance audit

A separate, later exploratory JUMP analysis used 771,769 treatment-level records and 614 numerical morphology features, with 652,578 compound records. The cross-source identity subset contained 5,000 shared compounds, each represented in source_6 and source_10. This experiment assessed whether the same compound could be distinguished from different compounds using morphological cosine similarity; it did not evaluate mechanism-of-action (MoA) labels. These results are distinct from the 65-record genetic perturbation pilot in Section IV-A and were not used to recalculate its plate-level statistical tests.

Source_6-only fitting of median imputation and standardization on 2,812 development compounds, followed by source_10 evaluation of 1,250 compound-disjoint held-out test identities, produced validation AUROC 0.6082 and test AUROC 0.6142. In a separate exploratory comparison with a fixed negative pairing, uncompressed cosine achieved AUROC 0.6102, versus 0.5979–0.6027 for PCA projections (32, 64, 128 and 256 dimensions); therefore PCA superiority is not claimed. One hundred alternative negative-pair shifts produced mean AUROC 0.6085 (SD 0.0040, range 0.5975–0.6192). These pair-sampling experiments do not substitute for independent biological replication, fully independent source panels, or MoA validation.

The evidence-grounded reasoning prototype produced three provisional compound-MoA research hypotheses and an auditable generate–critique–revise record. Separately, 11 CRISPR/ORF gene-pair morphology assessments underwent empirical-null and multiple-testing review: BRCA1–TP53 (CRISPR) yielded exploratory cosine similarity 0.4797 (empirical $p = 0.0185$), but its FDR-adjusted $q = 0.2037$; none of the eleven gene-pair tests met $q < 0.05$. One candidate was placed in a prioritized simulated human-review queue, without completed expert adjudication or verified biological discovery. These gene-pair findings must not be equated with the distinct nine gene–modality sign tests of Section IV-A.

Existing programmed governance checks passed 1,000/1,000 routing tests, 5/5 adversarial-policy tests, 11/11 evidence-integrity checks and 5/5 tampering tests. The initial five-case live-agent run had 2/5 execution errors; both failed cases subsequently received LLM responses in a new two-case Groq retest. Neither the rule-based pass rates nor response availability measure scientific correctness, independent agent superiority or real human intervention. The cause of the historical execution errors remains unknown. All additional results are observational and limited to the archived test cases.

E. Consolidated Experimental Results

Evaluation — result and interpretation
Plate-level pilot — 65 records; 4/5 CRISPR and 0/4 ORF significant
Compound identity — 614 features; AUROC 0.6142 (n=1,250); not MoA
Gene-pair evidence — 11 comparisons; none FDR-significant
Governance — 1,021/1,021 predefined checks passed
Live execution — initial 2/5 errors; 2/2 fresh retest responses

V. DISCUSSION

Three distinct findings emerge. First, the original 65-record genetic pilot yielded four of five FDR-significant CRISPR gene–modality tests and zero of four ORF tests, with BRCA1 showing the strongest adjusted evidence ($q = 0.004883$). Second, an independently scoped 614-feature cross-source compound-identity study produced held-out AUROC 0.6142 across 1,250 compound-disjoint test identities, indicating measurable but modest morphological consistency across the evaluated sources. Third, eleven gene-pair comparisons yielded no FDR-significant mechanistic inference; the strongest exploratory BRCA1–TP53 CRISPR pairing had $q = 0.2037$. These statistics

address different units, null hypotheses and biological claims, and cannot be pooled into a discovery or agent accuracy rate.

The pilot's absence of adjusted-significant ORF results does not imply lack of biological effects: modality, plate representation, multiple-testing correction and feature construction all affect detectability. In particular, the ten ORF BRCA1 records derive from only five represented plates, illustrating possible pseudoreplication. The 614-feature experiment is likewise a compound-identity retrieval task, not target identification or validated mechanism-of-action prediction. Its cross-source signal remains modest, and the cosine representation outperformed the tested PCA projections only in the reported exploratory comparison—not as a general statement about dimensionality reduction.

The central scientific implication is evidentiary separation. Morphology provides quantitative phenotype observations; gene–pathway links offer biological context; LLM-generated explanations remain provisional; and expert-confirmed mechanisms require independent experimental validation. This boundary is consistent with JUMP genetic perturbation mapping [8] and confounder-aware profiling [9], [11], [47], and it prevents attractive narratives from being mistaken for causal discovery.

On operational feasibility, the recorded governance suite passed all 1,021 predefined checks (1,000 routing, five adversarial, eleven integrity and five tampering). These checks establish behavior on the archived cases only; they are not a prospective safety guarantee. Initial live-agent execution failed for two of five cases, while fresh calls for both previously failed cases returned LLM responses. The historical failure cause was not recovered. Response availability, JSON correctness and programmed deferral do not establish scientific reasoning accuracy, real human review or agent superiority.

A decisive future evaluation should preregister experimental splits and agent prompts, freeze the retrieval corpus and model versions, measure uncertainty and correct evidence attribution, and compare statistics-only, retrieval-only, single-agent and multi-agent variants under identical case labels. Prospective biological verification and qualified human adjudication are required to measure genuine discovery utility. Until then, DISCOVERY-AGENT is best characterized as an exploratory, evidence-governed research workflow with documented feasibility and clear validation gaps.

VI. LIMITATIONS AND REPRODUCIBILITY

Despite providing exploratory evidence of perturbation-associated morphological responses and outlining an evidence-grounded autonomous discovery architecture, this study has several methodological, computational, and biological limitations that warrant careful consideration.

First, the empirical investigation is restricted to a small, selectively examined panel of genetic perturbations. The analysis encompasses five CRISPR-associated genes and four ORF-associated genes, limiting statistical power, biological diversity, and generalizability to the wider JUMP perturbation landscape. Although significant CRISPR-associated morphological signals were identified for AKT1, BRCA1, MTOR, and TP53, these findings cannot be generalized to untested genes, cellular environments, or intervention classes. Moreover, the analyzed perturbations were not selected through a prospectively registered sampling strategy. Consequently, the reported significance levels should be interpreted as exploratory rather than confirmatory evidence. This limitation is particularly relevant given recent findings that cellular perturbation models may exhibit considerable sensitivity to evaluation design, experimental context, and baseline selection [1], [26], [2].

Second, technical and biological sources of variation remain important potential influences on the observed results. High-content cellular measurements may be affected by plate-specific conditions, batch effects, cell-state heterogeneity, differences in perturbation efficiency, and preprocessing decisions. Recent advances in contrastive batch correction and confounder-aware morphological representation learning demonstrate the importance of separating biologically informative variation from experimental artifacts [9]–[11], [47]. Although the present analysis employs plate-level summaries and multiple-testing correction, these procedures do not independently establish freedom from confounding. Furthermore, the available records do not support a comprehensive assessment of cross-batch reproducibility, independent biological replication, or generalization across laboratory environments. Repeated observations associated with the same experimental plate must therefore not be treated as independent biological replicates.

Third, reproducibility is partial and varies across the research components. The 65-record genetic pilot has archived summary tables and figures but lacks all original upstream inputs, profile extraction steps and control/null-score transformations needed for complete re-execution. The later compound study has an analysis notebook, source-indexed Parquet data, saved splits and evaluation artifacts, yet a clean-environment rerun remains unverified. Live LLM inference further depends on available provider quotas, model versions and preserved prompts. Accordingly, integrity of the exported ZIP archive and correctness of saved metrics must not be confused with independent end-to-end reproduction.

Fourth, the reported morphological findings do not independently establish causal mechanisms, molecular targets, or functional gene interactions. Phenotypic similarity may motivate biological hypotheses but cannot replace molecular characterization or targeted experimental verification. Recent work combining Cell Painting with transcriptomics, proteomics, and chemical representations provides relevant pathways toward stronger mechanistic interpretation [19]–[21]. Likewise, network-based and AI-assisted drug-target identification research emphasizes the value of integrating multiple evidence sources [48], [49], [51]. Future investigations should therefore examine whether the identified signatures are reproducible across independent experiments and whether proposed mechanisms can be supported through orthogonal biological measurements.

Fifth, DISCOVERY-AGENT remains a proposed computational architecture rather than an experimentally validated autonomous discovery system. The current investigation does not establish improvements in mechanistic hypothesis accuracy, evidence-grounding reliability, compound–gene matching, agent collaboration, or biological discovery efficiency. Exploratory interactions with external language-model services are insufficient for comparative performance claims without frozen prompts, reproducible model configurations, independent evaluation labels, and controlled baselines. This distinction is particularly important because recent research has demonstrated that more sophisticated computational models do not necessarily outperform simpler alternatives [27], [2]. Future evaluation must therefore compare statistical-only analysis, conventional information retrieval, single-agent reasoning, and multi-agent configurations under consistent experimental conditions.

Sixth, evidence-grounded autonomy introduces challenges involving retrieval reliability, uncertainty calibration, and scientific accountability. Retrieved biological associations may be incomplete, contradictory, or context-dependent, while agreement between multiple AI agents may reflect shared assumptions rather than independent verification. Knowledge-graph-based hypothesis generation [41], [42], biomedical RAG [43], and uncertainty-aware human-in-the-loop validation [52] provide methodological foundations for addressing these concerns. Nevertheless, DISCOVERY-AGENT must undergo dedicated evaluation to determine whether its proposed verification mechanisms reduce unsupported hypotheses without excessively suppressing potentially valuable discoveries.

Beyond these technical limitations, the study raises important considerations regarding equitable participation in computational biological research. For research institutions in Pakistan and other countries across the Global South, large-scale experimental screening infrastructure, specialized imaging equipment, and sustained high-performance computing resources may present substantial practical barriers. Publicly accessible biological resources such as JUMP can offer an alternative entry point for computational investigation, allowing researchers to examine existing experimental evidence without reproducing the full infrastructure required for its initial generation [5]–[7].

However, open data alone does not eliminate disparities in scientific capacity. Access to computational resources, reliable research infrastructure, specialized training, curated biological knowledge, and opportunities for independent experimental validation remains essential. Consequently, the long-term relevance of DISCOVERY-AGENT to Pakistan and comparable research environments depends on whether its computational components can be made transparent, resource-efficient, reproducible, and compatible with locally available infrastructure. These potential benefits remain prospective and have not been established through a dedicated deployment or cost-effectiveness study.

Future research should therefore prioritize a publicly reproducible implementation, computationally efficient baseline models, auditable evidence retrieval, and collaborative validation involving institutions with complementary computational and experimental capabilities. Such an approach could create opportunities for researchers in Pakistan and the wider Global South to participate more actively in international computational biology through transparent reuse of public datasets and scientifically accountable AI-assisted hypothesis development. Importantly, these opportunities require empirical demonstration rather than assumptions about automatic technological accessibility.

Overall, the study's limitations define a clear validation roadmap. Establishing the scientific utility of DISCOVERY-AGENT will require broader perturbation coverage, independent biological replication, leakage-aware benchmarking, transparent source attribution, uncertainty calibration, and prospective experimental assessment. Addressing these requirements is essential before the proposed architecture can be considered a reliable autonomous biological discovery system.

VII. CONCLUSION

DISCOVERY-AGENT brings together three complementary exploratory evaluations of JUMP-derived cellular evidence. In the original 65-record CRISPR/ORF pilot, four CRISPR gene tests (AKT1, BRCA1, MTOR and TP53) met the adjusted significance threshold, whereas no evaluated ORF test did. Separately, a 614-feature cross-source compound-identity task achieved held-out AUROC 0.6142 on 1,250 compound-disjoint test identities. Eleven additional gene-pair morphological assessments yielded no FDR-significant mechanistic finding (minimum q = 0.2037). These results demonstrate measurable morphological structure while falling short of experimentally confirmed mechanisms or verified autonomous discovery.

The framework organizes provenance-aware phenotype analysis, literature and pathway context, provisional hypothesis generation, independent critique, uncertainty-sensitive triage and bounded human-review routing. Three hypothesis records were generated, critiqued and revised. Across archived governance tests, 1,021 of 1,021 predefined checks passed; however, the first five-case live-agent run had two execution failures, both of which returned responses on fresh retest. Such operational feasibility and policy compliance are not measures of biological truth, complete system safety, or superiority over statistical-only and single-agent baselines.

The main methodological contribution is therefore the explicit separation of observed phenotypic associations, cross-source identity signals, literature-derived context, machine-generated hypotheses, and expert-confirmed biological mechanisms. This evidentiary hierarchy is especially relevant to open-data research settings, including institutions in Pakistan and the Global South, but accessible datasets alone do not guarantee low-cost execution or experimentally validated benefit.

Next steps include publishing immutable data manifests and preprocessing dependencies, verifying a clean Kaggle run, using compound- and plate-aware prospective splits, and benchmarking retrieval-only, single-agent and multi-agent workflows with adjudicated evidence labels. Independent experimental perturbation or biochemical follow-up remains necessary before mechanistic claims can be confirmed.

DISCOVERY-AGENT should thus be interpreted as an auditable exploratory scientific-reasoning framework supported by multiple distinct computational assessments—not as a completed autonomous discovery system. Its value is a concrete evaluation pathway linking morphological data, conservative evidence governance and transparent human-directed validation.

REFERENCES

[1] G. Csendes et al., “Benchmarking foundation cell models for post-perturbation RNA-seq prediction,” BMC Genomics, 2025. https://doi.org/10.1186/s12864-025-11600-2

[2] C. Ahlmann-Eltze, W. Huber, and S. Anders, “Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines,” Nature Methods, 2025. https://doi.org/10.1038/s41592-025-02772-6

[3] “Machine learning to dissect perturbations in complex cellular systems,” Computational and Structural Biotechnology Journal, 2025. https://doi.org/10.1016/j.csbj.2025.02.028

[4] “Toward trustworthy virtual cells: a roadmap for perturbation-resolved, context-aware, and experimentally validated cell models,” Frontiers in Cell and Developmental Biology, 2026. https://doi.org/10.3389/fcell.2026.1900624

[5] The JUMP Cell Painting Consortium, “The Joint Undertaking for Morphological Profiling (JUMP) Consortium Datasets,” Zenodo, 2024. https://doi.org/10.5281/zenodo.12983164

[6] “Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations,” Nature Methods, 2024. https://doi.org/10.1038/s41592-024-02241-6

[7] E. Weisbart et al., “Cell Painting Gallery: an open resource for image-based profiling,” Nature Methods, vol. 21, pp. 1775–1777, 2024. https://doi.org/10.1038/s41592-024-02399-z

[8] S. N. Chandrasekaran et al., “Morphological map of under- and overexpression of genes in human cells,” Nature Methods, vol. 22, pp. 1742–1752, 2025. https://doi.org/10.1038/s41592-025-02753-9

[9] C. Yan et al., “Triple-effect correction for Cell Painting data with contrastive and domain-adversarial learning,” Nature Communications, 2025. https://doi.org/10.1038/s41467-025-62193-z

[10] V. Kim et al., “Self-supervision advances morphological profiling by unlocking powerful image representations,” Scientific Reports, 2025. https://doi.org/10.1038/s41598-025-88825-4

[11] G. Papanastasiou et al., “Confounder-aware foundation modeling for accurate phenotype profiling in cell imaging,” npj Imaging, 2025. https://doi.org/10.1038/s44303-025-00116-9

[12] E. von Coburg et al., “Cell Painting PLUS: expanding the multiplexing capacity of Cell Painting-based phenotypic profiling using iterative staining-elution cycles,” Nature Communications, 2025. https://doi.org/10.1038/s41467-025-58765-8

[13] M. A. Zietek et al., “Cell Painting in activated cells illuminates phenotypic dark space and drug mechanisms of action,” Nature Communications, 2026. https://doi.org/10.1038/s41467-026-76252-6

[14] Y. Tan et al., “Deciphering senescence-associated mechanisms through cell painting & transcriptomics,” npj Aging, 2026. https://doi.org/10.1038/s41514-026-00500-8

[15] M. Sanchez et al., “Large scale compound selection guided by cell painting reveals activity cliffs and functional relationships,” Communications Biology, 2026. https://doi.org/10.1038/s42003-025-09500-y

[16] S. Sivanandan et al., “A pooled Cell Painting CRISPR screening platform enables de novo inference of gene function by self-supervised deep learning,” Nature Communications, 2026. https://doi.org/10.1038/s41467-025-66778-6

[17] D. Sun et al., “CPHNet: a novel pipeline for anti-HAPE drug screening via deep learning-based Cell Painting scoring,” 2025. https://doi.org/10.1186/s12931-025-03173-1

[18] J. D. Ewald et al., “Cell Painting for cytotoxicity and mode-of-action analysis in primary human hepatocytes,” Cell Systems, 2026. https://doi.org/10.1016/j.cels.2026.101566

[19] S. V. Ha et al., “Cross modality learning of cell painting and transcriptomics data improves mechanism of action clustering and bioactivity modelling,” Scientific Reports, 2025. https://doi.org/10.1038/s41598-025-05914-0

[20] Y. Yu et al., “Integrating chemical structures as treatments improves representations of microscopy images for morphological profiling,” PLOS Computational Biology, 2026. https://doi.org/10.1371/journal.pcbi.1014483

[21] C. Johansson et al., “Cell painting and thermal proteome profiling for inference of drug targets and mechanism of action,” Molecular Systems Biology, 2026. https://doi.org/10.1038/s44320-026-00214-9

[22] S. Baek, K. Song, and I. Lee, “Single-cell foundation models: bringing artificial intelligence into cell biology,” Experimental & Molecular Medicine, 2025. https://doi.org/10.1038/s12276-025-01547-5

[23] Y. Wang et al., “scREPA: Predicting single-cell perturbation responses with cycle-consistent representation alignment,” Computational Biology and Chemistry, 2025. https://doi.org/10.1016/j.compbiolchem.2025.108712

[24] E. Cole et al., “Foundation Models Improve Perturbation Response Prediction,” bioRxiv, 2026, preprint. https://doi.org/10.64898/2026.02.18.706454

[25] E. Hasanaj et al., “Multimodal Benchmarking of Foundation Model Representations for Cellular Perturbation Response Prediction,” bioRxiv, 2025, preprint. https://doi.org/10.1101/2025.06.26.661186

[26] M. Vollenweider and P. Bühlmann, “Signal, Bounds, and Baselines: Principles for Evaluating Virtual Cell Perturbation Models,” bioRxiv, 2026, preprint. https://doi.org/10.64898/2026.04.20.719650

[27] G. Palla et al., “Tabular Foundation Models Are Competitive Cellular Perturbation Predictors Across Biological Scales,” bioRxiv, 2026, preprint. https://doi.org/10.64898/2026.06.28.735106

[28] R. Peng et al., “A generative framework for predicting cellular morphological and transcriptomic perturbation responses,” Cell Reports Methods, 2026. https://doi.org/10.1016/j.crmeth.2026.101459

[29] A. E. Ghareeb et al., “A multi-agent system for automating scientific discovery,” Nature, vol. 655, pp. 497–505, 2026. https://doi.org/10.1038/s41586-026-10652-y

[30] C. Qi et al., “Artificial Intelligence agents for biological research: A survey,” Briefings in Bioinformatics, 2026. https://doi.org/10.1093/bib/bbag075

[31] F. Trost et al., “An agentic framework for autonomous scientific discovery in cancer pathology,” Nature Medicine, 2026. https://doi.org/10.1038/s41591-026-04357-y

[32] D. L. Huynh and S. Seal, “AI agents in drug discovery: applications and case studies,” Drug Discovery Today, 2026. https://doi.org/10.1016/j.drudis.2026.104650

[33] “Streamline automated biomedical discoveries with agentic bioinformatics,” Briefings in Bioinformatics, 2025. https://doi.org/10.1093/bib/bbaf505

[34] “The next paradigm in bioinformatics: a review of multi-agent systems and foundational models for end-to-end scientific discovery,” Briefings in Bioinformatics, 2026. https://doi.org/10.1093/bib/bbag245

[35] “Bridging LLM-based agents and workflow languages for automated multi-perspective analysis of scRNA-seq data,” Bioinformatics Advances, 2026. https://doi.org/10.1093/bioadv/vbag301

[36] “Large language model agents for biological intelligence across genomics, proteomics, spatial biology, and biomedicine,” Briefings in Bioinformatics, 2026. https://doi.org/10.1093/bib/bbag110

[37] “Building MCP-native hierarchical AI scientist ecosystems: a perspective on scaling multi-agent scientific discovery,”

Frontiers in Artificial Intelligence, 2026. https://doi.org/10.3389/frai.2026.1820375

[38] B. Li et al., "Agentic AI and the rise of in silico team science in biomedical research," Nature Biotechnology, 2026. https://doi.org/10.1038/s41587-026-03035-1

[39] "Autonomous Self-Evolving Research on Biomedical Data: The DREAM Paradigm," Advanced Science, 2025. https://doi.org/10.1002/advs.202417066

[40] "AI-Generated Hypotheses and the Emergence of Autonomous Scientific Discovery," ACS Materials Letters, 2026. https://doi.org/10.1021/acsmaterialslett.6c00224

[41] M. Yang, K. Dong, K. Lim, and J. Liu, "THGAgents: Traceable Biomedical Hypothesis Generation via Dynamic Causal Reasoning," in Proc. IJCAI, 2026. https://doi.org/10.24963/ijcai.2026/776

[42] A. Borrego et al., "Research hypothesis generation over scientific knowledge graphs," Knowledge-Based Systems, 2025. https://doi.org/10.1016/j.knosys.2025.113280

[43] "BioRAGent: Natural language biomedical querying with retrieval-augmented multiagent systems," Briefings in Bioinformatics, 2025. https://doi.org/10.1093/bib/bbaf539

[44] "Graph Retrieval-Augmented Generation: A Survey," ACM Transactions on Information Systems, 2026. https://doi.org/10.1145/3777378

[45] "Federated knowledge retrieval elevates large language model performance on biomedical benchmarks," GigaScience, 2026. https://doi.org/10.1093/gigascience/giag007

[46] M. Yuan, T. Wang, and Y. Qian, "AdaKG-RAG: Adaptive KG-guided retrieval-augmented generation with hypothesis-and-verification retrieval for multi-hop question answering," Journal of King Saud University – Computer and Information Sciences, 2026. https://doi.org/10.1007/s44443-026-01028-3

[47] "Adversarial batch representation augmentation for batch correction in high-content cellular screening," Knowledge-Based Systems, 2026. https://doi.org/10.1016/j.knosys.2026.115829

[48] T. Koutsandreas et al., "Network-Based Approaches for Drug Target Identification," Annual Review of Biomedical Data Science, 2025. https://doi.org/10.1146/annurev-biodatasci-101424-120950

[49] F. W. Pun et al., "Target identification and assessment in the era of AI," Nature Reviews Drug Discovery, 2026. https://doi.org/10.1038/s41573-026-01412-8

[50] "Predicting the effects of drugs and unveiling their mechanisms of action using an interpretable pharmacodynamic mechanism knowledge graph (IPM-KG)," Computers in Biology and Medicine, 2025. https://doi.org/10.1016/j.compbiomed.2024.109419

[51] "Intelligence on the graph: Graph neural networks for mechanistic drug target discovery," Journal of Pharmaceutical Analysis, 2026. https://doi.org/10.1016/j.jpha.2025.101508

[52] "Conformal validation: A deferral policy using uncertainty quantification with a human-in-the-loop for model validation," Machine Learning with Applications, 2025. https://doi.org/10.1016/j.mlwa.2025.100733

[53] X. Kuznetsova, L. Kuznetsova, E. Shabunina, I. Sergeev, and I. Malyshev, "Retrieval-Based Evaluation of Cell Painting Feature Spaces Reveals Differences in the Preservation of Biologically Meaningful Phenotypic Similarity," International Journal of Molecular Sciences, vol. 27, no. 13, Art. no. 5727, 2026. https://doi.org/10.3390/ijms27135727