

Auditing Retrieval-Augmented LLM Hypotheses for Longitudinal Cell Painting Morphology

Gilchan Park, Guang Zhao, Byung-Jun Yoon, Shinjae Yoo
 Brookhaven National Laboratory
 Upton, New York, USA
 {gpark,gzhao,byoon,sjyoo}@bnl.gov

Abstract

High-content morphological profiling (Cell Painting) yields sensitive, high-dimensional signatures of cellular state, but translating longitudinal morphology trajectories into interpretable biology remains difficult—especially for weak, chronic perturbations such as low-dose-rate ionizing radiation. Large language models (LLMs) can synthesize heterogeneous evidence into narratives, yet scientific use requires quantitative auditing: fluent outputs may be ungrounded or inconsistent with measured morphology.

We present an evaluation-first, retrieval-augmented interpretation framework for longitudinal Cell Painting morphology. To demonstrate the framework’s utility in a highly challenging real-world scenario characterized by weak signals and sparse reference data, we apply it to a 9-week RPE-1 time course spanning five dose rates (0.003–6.0 mGy/hr). We compute week-matched treated–control morphology deltas and construct evidence-rich prompt payloads that bind (i) quantitative deltas to (ii) retrieved perturbation neighbors, pathway context, and literature snippets via stable evidence identifiers. An LLM generates structured controlled-vocabulary hypotheses at the week×dose level, which are then integrated hierarchically into dose-time phases and global summaries while preserving evidence provenance.

We operationalize auditing with two quantitative tests: V1 citation validity, which verifies that every cited evidence identifier is present in the prompt payload, and V2 proxy-based morphology compatibility, which measures whether predicted process labels are consistent with the top changed morphology features under an explicit proxy table. In the audited run, V1 detected no invalid evidence references, and V2 showed non-trivial morphology compatibility that increased with perturbation strength. At week×dose resolution, V2 also showed a modest positive association with an LLM-external morphology drift summary computed from the same Cell Painting data distribution. Crucially, this auditable pipeline distilled complex trajectories into falsifiable hypotheses, including a distinct adaptive phenotype associated with metabolic reprogramming and proteostatic stress at lower dose rates (0.003–0.3 mGy/hr). Limitations include proxy-based evaluation granularity and the absence of ground-truth mechanism labels. The source code and datasets for this study are publicly available at <https://github.com/cellpainting-llm-auditor>.

CCS Concepts

• **Applied computing** → **Bioinformatics**.

Keywords

Cell Painting, morphological profiling, longitudinal phenotyping, low-dose ionizing radiation (LDIR), LDR, Large Language Models

(LLMs), retrieval-augmented generation, LLM auditing, hypothesis generation

1 Introduction

High-content imaging assays provide a sensitive readout of how perturbations reshape cellular state. Cell Painting–style morphological profiling converts multiplexed fluorescence images into quantitative single-cell feature vectors spanning size, shape, intensity, and texture [3]. Repeated measurement enables longitudinal trajectories, but mechanistic interpretation is particularly challenging for weak, chronic perturbations such as low-dose and low-dose-rate ionizing radiation, where responses can be multifactorial and context dependent [22, 31].

The interpretability bottleneck is largely downstream of feature extraction. Morphological profiles are high-dimensional and sensitive to technical variation, often requiring robust normalization and aggregation before comparisons are meaningful across plates and timepoints [4, 5]. Even after preprocessing, translating a trajectory in morphology space into biological hypotheses typically remains manual. Public reference resources partially mitigate this gap via “profile matching,” where new signatures are compared to large perturbation libraries (e.g., JUMP) to suggest candidate gene programs or pathways [5]. However, similarity alone rarely yields an end-to-end explanation—especially in longitudinal settings where dominant processes may shift by dose or phase.

Large language models (LLMs) offer an appealing interface for synthesizing heterogeneous evidence, but scientific deployment requires auditing: fluent narratives can be unfaithful to evidence and may include fabricated or incorrect citations [11, 28]. Retrieval-augmented generation (RAG) improves provenance by coupling generation to external evidence stores [15], and curated resources such as Reactome provide standardized pathway context [20].

In this work, we introduce an LLM-centered, retrieval-augmented interpretation framework for longitudinal Cell Painting morphology that is evaluation-first by design. To demonstrate the practical utility of this framework in a highly challenging real-world scenario, we apply it to the longitudinal profiling of chronic low-dose-rate (LDR) ionizing radiation—a domain where morphological signals are notoriously weak and established reference mechanisms are scarce. Rather than treating LLM narratives as conclusions, we treat them as auditable scientific artifacts: structured hypotheses grounded in (i) measured morphology deltas, and (ii) retrieved external evidence (perturbation neighbors, pathway context, literature snippets).

Specifically, we operationalize auditing through two rigorous quantitative tests. First, we enforce citation integrity (V1) by verifying that every piece of evidence cited by the LLM strictly maps

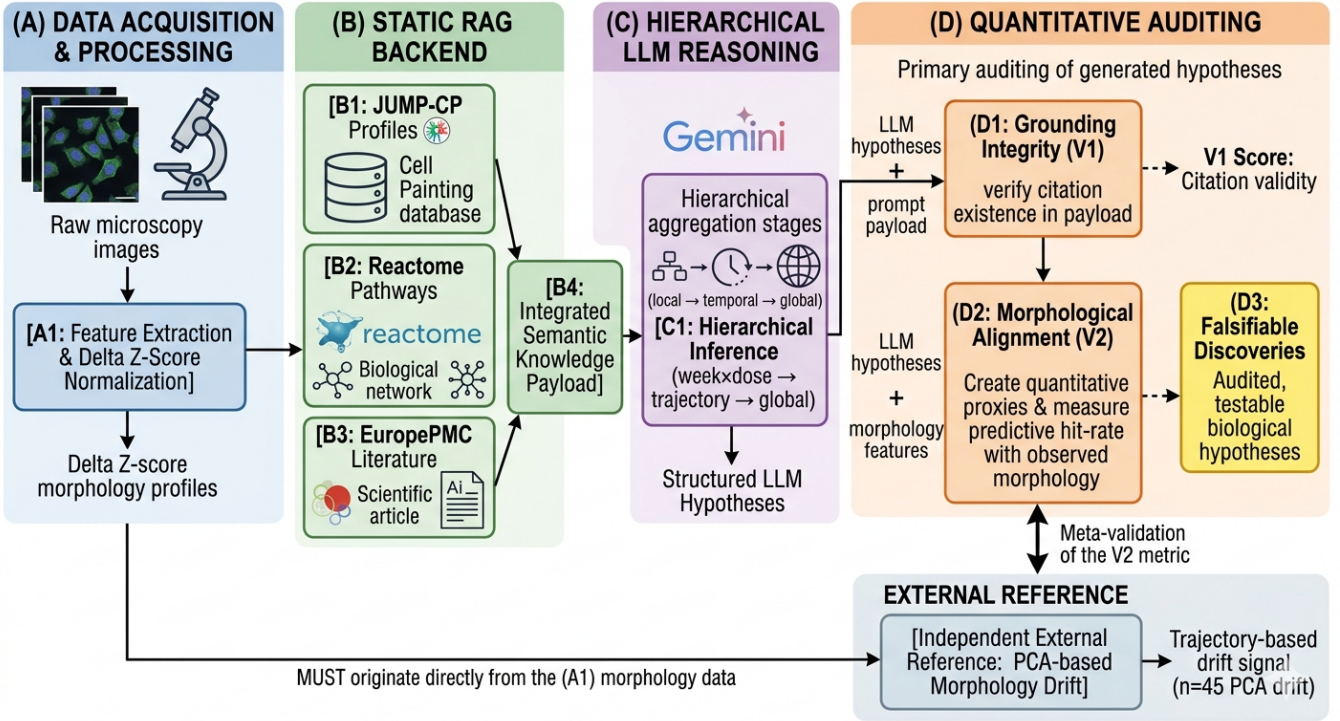


Figure 1: Architectural overview of the evaluation-first LLM-driven morphological reasoning and validation pipeline. The framework consists of four integrated modules: (A) Data Acquisition & Processing, where raw microscopy images undergo feature extraction and delta Z-score normalization to yield quantitative morphology profiles; (B) Static RAG Backend, which constructs an integrated semantic knowledge payload by synthesizing morphological reference profiles (JUMP-CP), mechanistic biological pathways (Reactome), and scientific literature contexts (EuropePMC); (C) Hierarchical LLM Reasoning, employing the Gemini large language model to iteratively consolidate local condition-specific inferences (week × dose) into temporal trajectories and unified global hypotheses; and (D) Quantitative Auditing, the primary evaluation module where hypotheses are rigorously anchored against grounding integrity (V1: citation validity) and morphological alignment (V2: quantitative predictive hit-rate) to produce audited, falsifiable biological discoveries. Furthermore, the V2 auditing metric is meta-validated against an independent external reference of trajectory-based PCA morphology drift. (Note: The visual representation of this diagram was generated with the assistance of the Gemini.)

back to the provided prompt payload, effectively eliminating reference hallucinations. Second, we assess biological fidelity via a proxy-based alignment metric (V2), which evaluates the consistency between the LLM’s predicted biological processes and the most prominent morphological feature shifts. To further substantiate these findings, we establish an independent, morphology-only PCA drift summary. This drift metric serves as a crucial secondary quantitative anchor, ensuring that the semantic trajectories proposed by the LLM are fundamentally grounded in the principal axes of phenotypic variation observed in the raw data (Figure 1).

Contributions:

- **Evaluation-first auditing for LLM interpretation of morphology:** Formal validation of hypothesis grounding (V1) and quantitative proxy alignment (V2), designed specifically for settings lacking ground-truth mechanism labels.
- **Traceable retrieval-augmented prompting:** A robust prompt payload schema that binds quantitative morphology

deltas to retrieved perturbation and literature evidence via stable IDs, enabling automatic citation verification.

- **Hierarchical reasoning with provenance:** A multi-step synthesis approach (dose-time and global summaries) that preserves longitudinal structure and reliably propagates evidence references, preventing the loss of context typical of flattened week × dose outputs.

2 Related Work

2.1 Cell Painting and morphological profiling

Cell Painting is a widely adopted foundation for image-based morphological profiling, yielding multi-channel fluorescence images that can be summarized into quantitative single-cell feature vectors [3]. A recurring theme in the literature is that interpretability is limited less by feature extraction than by downstream processing and contextualization: profiles often require robust normalization,

aggregation, and careful similarity computation before they can be compared across batches or experimental conditions [5].

Recently, large public reference resources have emphasized both the promise and complexity of morphological profiling. The JUMP Cell Painting Consortium released a benchmark-scale dataset spanning matched chemical and genetic perturbations to support similarity evaluation and representation learning [5]. While early approaches relied on classical image analysis pipelines, modern efforts increasingly leverage deep representation learning—such as weakly supervised learning and vision transformers—to extract more robust phenotypic embeddings [4, 5]. Such resources enable “profile matching” as a practical mechanism-discovery heuristic, but nearest-neighbor similarity typically requires additional pathway and literature context. This remains especially challenging for longitudinal settings where phenotypes evolve dynamically over time.

2.2 LLMs, RAG systems, and biological interpretation

Biomedical LLMs have become central tools for text mining and synthesis, building on domain-adapted pretraining (e.g., BioBERT, PubMedBERT) and subsequent generative models tailored for biomedical knowledge tasks [8, 14, 17]. More recently, state-of-the-art models like Med-PaLM have demonstrated expert-level performance on medical and biological reasoning benchmarks, signaling a shift toward highly capable, domain-specific reasoning engines [29].

In parallel, retrieval-augmented generation (RAG) couples a generator with an evidence store to improve provenance and factuality for knowledge-intensive generation [15]. In biology, curated resources such as Reactome provide standardized pathway definitions and gene–pathway mappings that are well suited for evidence-aware interpretation pipelines [20]. The integration of RAG into scientific workflows is now moving beyond simple question-answering toward autonomous hypothesis generation and complex data interpretation [2].

2.3 Grounding, evaluation, and interpretation auditing

Despite their capabilities, neural generation in scientific settings can be fluent yet unfaithful to underlying evidence, motivating explicit grounding constraints and automated checks [1, 19]. Prior work proposes faithfulness evaluations and truthfulness benchmarks (e.g., TruthfulQA) [16] and scientific claim verification benchmarks (e.g., SciFact) [32]. More broadly, documentation and auditing frameworks (e.g., model cards, datasheets) motivate treating AI artifacts as objects requiring structured reporting and evaluation [7, 21].

To the best of our knowledge, no existing computational framework natively integrates longitudinal Cell Painting trajectories with RAG-driven LLM synthesis and quantitative auditing. Traditional morphological profiling largely relies on single-time-point proximity matching (e.g., using JUMP-CP benchmarks), which limits its ability to capture weak, continuous perturbations unfolding over multiple weeks. Conversely, standard RAG applications in biology primarily operate on textual data and lack direct grounding in quantitative, image-derived morphological changes.

Our work bridges this gap by introducing an *evaluation-first auditing* framework for LLM-generated hypotheses in longitudinal morphology. Rather than assuming that retrieval augmentation alone guarantees correctness, we formalize measurable constraints—such as grounding integrity—and introduce quantitative alignment tests, including proxy-based morphology consistency. Crucially, these LLM-centric evaluation metrics are anchored to an independent morphology-only reference used strictly for validation, enabling a new paradigm of auditable morphological reasoning where direct empirical comparison to existing approaches remains challenging.

3 Methods

3.1 Experimental design and dataset

We studied long-term cellular responses to chronic low-dose ionizing radiation in RPE-1 (human retinal pigment epithelial) cells using the Cell Painting assay [3]. Cells were exposed at five dose rates (mGy/hr): 0.003, 0.03, 0.3, 3.0, and 6.0, and profiled longitudinally from week 1 through week 9 (9 weeks total; Table A1).

Paired treated–control design. For each week, the treated plate (designated as Plate 3) was paired with a same-week control plate (Plate C). Within each week w and dose d , we analyzed four treated replicate wells (180 treated wells total). Control wells were defined on Plate C (2 wells in week 1; 4 wells in weeks 2–9; 34 control wells total), yielding 214 wells analyzed overall (Table A1).

Cell-level file structure. Cell segmentation and feature extraction produced per-image tables of single-cell features. For each well, images were acquired in $f \in \{1, \dots, 9\}$ fields of view (FOVs) and $p \in \{1, \dots, 15\}$ focal planes (135 images per well nominal). Across the study, 28,886 cell-level tables were analyzed (expected 28,890; four files missing), comprising 6,375,699 single-cell observations (Table A1). Single-cell segmentation was performed using the CellSAM model [18] to extract instance-level masks prior to morphological feature computation.

3.2 Morphology preprocessing pipeline (Steps 1–3)

The preprocessing pipeline converts cell-level observations into well-level treated–control deltas suitable for retrieval and LLM prompting.

3.2.1 Metadata construction (Step 1). We parse the directory structure and file naming convention to map each cell-level file to a well identifier and experimental context (w, d , condition) under the paired Plate 3/Plate C design. The result is stored as a metadata index table.

3.2.2 Cell-level to well-level aggregation (Step 2). For each well, we pool all single-cell observations across its available FOV×plane files and compute robust summary statistics for each base morphology feature: median, median absolute deviation (MAD), 10th percentile, and 90th percentile. We exclude geometry/QC-like columns orientation, centroid_x, and centroid_y during aggregation. For the nine base features (area, perimeter, eccentricity, solidity, mean_intensity, glcm_contrast, glcm_correlation, glcm_energy, glcm_homogeneity), this yields 36 well-level summary features.

3.2.3 Robust within-week normalization and treated-control deltas (Step 3). Let $x_{i,f}$ denote the aggregated value for well i and feature f , and let $w(i)$ denote the week of well i . We perform robust within-week normalization using the per-week feature median and MAD computed across *all* wells (treated and control) in the same week. Letting $W(i)$ denote the set of all wells in week $w(i)$, and with $\epsilon = 10^{-6}$, the normalized value is

$$z_{i,f} = \frac{x_{i,f} - \text{median}_{j \in W(i)} x_{j,f}}{\text{MAD}_{j \in W(i)} x_{j,f} + \epsilon}. \quad (1)$$

We then define the week-specific control baseline vector as the feature-wise median of normalized control wells C_w in week w . The treated-control delta for well i is

$$\Delta_{i,f} = z_{i,f} - \text{median}_{j \in C_{w(i)}} z_{j,f}. \quad (2)$$

To ensure stable baselines, we require at least two control wells per week and record control coverage diagnostics. Dataset scale and preprocessing exclusions are summarized in Table A1.

3.3 LLM-external morphology drift summary (Step 4)

We compute an LLM-external summary of morphology drift over time for validation-oriented sanity checks only. This module is independent of the retrieval and generation stack, but it is still derived from the same Cell Painting data distribution and is therefore not treated as independent biological validation.

3.3.1 PCA fitting and drift definition. We fit PCA on the normalized well profiles (z) restricted to control wells (Plate C) across all weeks. We use $n_{\text{PC}} = 12$ components and apply whitening. We then project each delta profile Δ_i into this PCA space and define drift magnitude as the Euclidean norm of the whitened PCA score vector:

$$D_i = \|\text{PCA}_{\text{whiten}}(\Delta_i)\|_2. \quad (3)$$

For each (w, d) , we aggregate drift by averaging D_i across the four treated replicate wells.

3.3.2 Week×dose and dose-level summaries. We summarize drift at two resolutions. First, for each (w, d) we retain the week×dose mean drift as a higher-resolution morphology-only summary. Second, for each dose d , we summarize the week-by-week drift trajectory using (i) a linear slope computed by least-squares regression of mean drift against week index, and (ii) an area-under-the-curve (AUC) computed by trapezoidal integration over weeks. These summaries are used only as LLM-external quantitative anchors for sanity checks on score behavior.

3.4 Retrieval-augmented prompt construction (Steps 5–6)

We construct a structured evidence payload for each (w, d) to support grounded, traceable LLM hypothesis generation.

3.4.1 Reduced morphology neighbor index. We build a reduced-feature index over a public Cell Painting perturbation reference (JUMP). We compute a 9-dimensional embedding corresponding to the base morphology feature set above by mapping each feature to matching Cell Painting columns and averaging across available channels/scales. Reference embeddings are z-scored using mean

and standard deviation with $\epsilon = 10^{-8}$. The index uses exact L2 neighbor search (FAISS when available), otherwise a scikit-learn nearest-neighbor backend with Euclidean distance [12, 25].

3.4.2 Neighbor search and selection. For each (w, d) observation, we query the reference index with $k = 100$ candidate neighbors and retain up to 10 for prompting, enforcing that at least three neighbors correspond to gene perturbations when possible. Neighbor records are assigned stable evidence IDs of the form `jump:<jump_doc_id>`.

3.4.3 Reactome pathway retrieval. We construct a retrieval index over Reactome pathway descriptions using TF-IDF (1–2 grams; cosine similarity; min_df=2; max_features=250000) [20]. For each (w, d) payload, we aggregate enriched pathways for Entrez genes associated with the candidate neighbors and retain the top 12 pathways by frequency. For dose-time and global aggregation (Step 8), pathway references can be cited using `path:<pathway_id_or_name>` and are validated against the payload; for week×dose generation (Step 7), enforceable retrieval IDs are limited to `jump:` and `lit:` (see below).

3.4.4 Literature snippets and quantitative observations. When enabled, we retrieve brief literature snippets per prompt via the Europe PMC REST API [6], using radiation-focused queries augmented with the cell line and additional terms (dose rate, protracted, long-term). We retrieve up to 10 hits per query and optionally pull and truncate selected full-text sections when available (sections methods, results, discussion, supplementary; section character cap 5000). Each snippet is assigned a stable ID `lit:<query_idx>:<lit_id>`. Each payload also includes a quantitative observation block with a unique ID `obs_w<week>_d<dose>` and the top changed morphology features by absolute mean delta.

3.4.5 Prompt payload schema and traceability. Each JSONL prompt record stores: (i) the observation block (obs ID and top-changed features), (ii) selected neighbors (with `jump:` IDs), (iii) literature snippets (with `lit:` IDs), and (iv) pathway evidence (pathway IDs). These evidence identifiers form the reference set used for grounding validation (V1).

3.5 Structured LLM hypothesis generation (Step 7)

We run an LLM reasoner per (w, d) prompt with strict structured-output constraints.

3.5.1 Model and decoding. We use Google’s Gemini model via Vertex AI [30] with temperature 0.2, thinking level HIGH, maximum output length 8192 tokens, and a prompt length cap of 300,000 characters. The runner retries up to three times on schema or grounding failures. The specific model revision served by the endpoint was `gemini-3-pro-preview`.

3.5.2 JSON-only schema and controlled vocabulary. The model is instructed to output JSON only, conforming to a schema with one or more hypotheses per prompt. Each hypothesis includes: `hypothesis_id`, `biological_process`, `summary`, `quant_evidence_refs`, `retrieval_refs`, `testable_predictions`, `confidence`, and a categorical

process_label constrained to an 11-label controlled vocabulary (Table A2). Schema validation is enforced programmatically.

3.5.3 Grounding enforcement. For week×dose hypotheses, grounding validation requires that all quant_evidence_refs are of the form obs:<obs_id> and match the prompt’s observation ID, and that all retrieval_refs are of the form jump:<jump_id> or lit:<lit_id> and appear in the payload. We also require that each hypothesis cites at least one observation and one retrieved item.

3.6 Hierarchical integration across scales (Step 8)

Longitudinal interpretation benefits from hierarchical reasoning that preserves temporal structure.

3.6.1 Dose-time aggregation. We construct dose-time prompts by grouping week×dose outputs across weeks for each dose and adding a timepoint catalog that retains per-week observation IDs, selected neighbors, literature snippets, and pathway evidence. A second LLM pass produces phase-level summaries (e.g., early/mid/late) with evidence references that must cite at least one observation and one retrieved item per phase.

3.6.2 Why week × dose first (and why not flat aggregation). Week×dose reasoning is performed first because it is the smallest unit in which (i) quantitative deltas $\Delta_{i,f}$ are directly defined against the matched weekly control baseline, and (ii) retrieval evidence (nearest neighbors, pathway context, literature snippets) is conditioned on the same quantitative signature. Aggregating all timepoints into a single prompt would mix heterogeneous temporal regimes and can obscure non-monotonic dynamics: feature directions can cancel when averaged across weeks, phase-specific patterns can be lost, and trends can reverse when marginalizing over time (a Simpson’s paradox). Flat prompts also dilute provenance by coupling a large evidence pool to a small number of claims, increasing ambiguity about which observations support which statements.

To preserve traceability, we keep evidence identifiers stable across scales. Each week×dose prompt assigns a unique observation ID (e.g., obs:obs_w{week}_d{dose}) and retains retrieval IDs (jump:, lit:, and Reactome IDs). Dose_time prompts embed a per-week catalog of these IDs; global prompts embed the per-dose catalogs. Downstream, V1 grounding checks operate on these identifier sets to verify that every phase- or global-level statement cites concrete week-level measurements and retrieved evidence.

3.6.3 Global summarization. We further summarize across doses by constructing a global prompt that preserves the dose-time evidence catalog and produces a small set of global hypotheses with evidence references. Hierarchical reasoning avoids flattening heterogeneous timepoints and enables phase-specific claims while maintaining provenance.

3.7 Validation framework

We distinguish (i) measured morphology quantities (Δ , drift), (ii) retrieval evidence (neighbors, snippets, pathways), and (iii) LLM-generated hypotheses. We quantify two complementary validation layers.

3.7.1 V1: citation validity (grounding integrity). V1 checks citation integrity: do evidence references in an LLM hypothesis correspond to evidence IDs present in the associated prompt payload? Let $O(p)$ be the observation ID set for payload p and $\mathcal{E}(p)$ the set of retrieval IDs (jump: and lit: for week×dose). For a hypothesis h with reference sets $Q(h)$ (quantitative) and $\mathcal{R}(h)$ (retrieval), we define

$$V1(h, p) = \mathbb{I}\left[Q(h) \subseteq O(p) \wedge \mathcal{R}(h) \subseteq \mathcal{E}(p) \wedge |Q(h)| > 0 \wedge |\mathcal{R}(h)| > 0\right]. \quad (4)$$

V1 therefore evaluates identifier-level provenance validity within a closed evidence set. It does not test whether a cited item semantically entails or fully supports a claim.

3.7.2 V2: proxy-based process–morphology compatibility. V2 tests whether a predicted process_label is compatible with the prompt’s quantitative morphology signature. To operationalize this comparison, we use an explicit proxy mapping between the controlled vocabulary of biological processes and expected directional shifts in core morphological features (Table A6). Because morphological profiles are highly multiplexed, we restricted our proxy definitions to well-established, macroscopic cellular hallmarks. The mapping is intentionally simplified, but it is fully specified in the manuscript and programmatically applied, making the evaluation transparent and reproducible.

For instance, cellular senescence is classically characterized by cellular hypertrophy (significantly increased area and perimeter) alongside increased cytoplasmic granularity and vacuolization (reflected as increased glcm_contrast and decreased glcm_energy) [10]. Conversely, apoptosis and cell death typically manifest as severe cell shrinkage (decreased area) and chromatin condensation or membrane blebbing, which drive high textural contrast [13]. Furthermore, cell-cycle arrest frequently leads to continued macromolecular synthesis without division, resulting in increased overall cell size [23]. Processes such as oxidative stress and DNA damage response, while lacking a uniform macroscopic size shift, frequently induce cytoskeletal remodeling and nuclear texture alterations identifiable via GLCM metrics [26].

We emphasize that this predefined proxy mapping is not intended as a definitive single-cell mechanism classifier or a ground-truth benchmark. Rather, it serves as a conservative, literature-grounded consistency check designed to penalize biologically contradictory LLM outputs (e.g., an LLM predicting “apoptosis” when the observation payload clearly indicates massive cell enlargement).

Formally, for each label ℓ , let $\mathcal{P}_\ell$ be its defined proxy set over the reduced feature vocabulary. Let $\mathcal{T}$ denote the set of top features provided in the prompt (ranked by $|\Delta|$). We define Hit@10 as the fraction of top features belonging to the proxy set:

$$\text{Hit@10}(\ell) = \frac{|\mathcal{T} \cap \mathcal{P}_\ell|}{|\mathcal{T}|}. \quad (5)$$

We also compute a magnitude-aware weighted proxy score as the fraction of total absolute delta mass in $\mathcal{T}$ explained by proxy hits:

$$WP(\ell) = \frac{\sum_{f \in \mathcal{T} \cap \mathcal{P}_\ell} |\Delta_f|}{\sum_{f \in \mathcal{T}} |\Delta_f|}. \quad (6)$$

3.7.3 Consistency with the drift summary. To relate LLM-centric validation to an LLM-external morphology-only signal, we perform two correlation analyses. First, at week $\times$ dose resolution we aggregate per-hypothesis V2 scores to the observation level by mean (with median and max summaries reported in supplementary diagnostics) and correlate these summaries with the corresponding week $\times$ dose drift mean from Step 4. Second, we compute a coarser dose-level comparison using per-dose V2 summaries and dose-level drift summaries (mean drift, slope, and AUC). Pearson and Spearman correlations are reported with nonparametric bootstrap uncertainty (10,000 resamples). These analyses are intended as sanity checks on score behavior rather than as independent biological validation.

3.7.4 Supplementary diagnostics. Additional quantitative and retrieval diagnostics referenced throughout the paper are provided in Supplementary Materials in Appendix.

4 Results

4.1 Grounding integrity (V1)

Across the full longitudinal design (9 weeks $\times$ 5 dose rates), the week $\times$ dose LLM reasoning stage produced 45/45 successful prompt records (0 failures), yielding 136 total hypotheses. Under the implemented citation-integrity check, no grounding failures were detected: 0 invalid evidence-reference cases and 0 records missing grounding statistics.

V1 grounding validity was perfect under strict grounding criteria (Methods, Eq. 4). Both the mean observation-reference validity rate and the mean retrieval-reference validity rate were 1.0. Furthermore, all hypotheses contained at least one observation reference and one retrieval reference (coverage rate = 1.0). Aggregate V1 metrics are summarized in Table 3; additional reference-count diagnostics are provided in Table A5 and Figure A2. By construction, these V1 results establish provenance validity within the closed prompt evidence set; they do not by themselves show that a cited item semantically supports a claim.

4.2 Representative auditable hypotheses

To make the audit-first outputs concrete, we highlight one representative week $\times$ dose hypothesis alongside its supporting quantitative signature. Table 1 displays the input prompt payload for week 4 at 3.0 mGy/hr, which reports the top- $|\Delta|$ morphology deltas (group-mean ΔZ ; values rounded to three decimals).

Based on this signature and the retrieved context, the LLM generated hypothesis `hyp_w4_d3.0_1`, classifying the state as `senescence_like` with high confidence (Table 2). We present this example as an auditable biological hypothesis: its cited evidence is traceable, and its declared label can be checked for morphology-compatibility under V2.

Table 1: Quantitative prompt payload (top morphology deltas) for week 4 at 3.0 mGy/hr.

Feature	ΔZ
mean_intensity_median	+2.145
glcm_energy_median	-1.786
solidity_median	+1.364
glcm_homogeneity_median	-1.204
area_median	+1.007
glcm_correlation_median	-0.756
eccentricity_median	+0.621
glcm_contrast_median	-0.268

Table 2: Representative week $\times$ dose hypothesis generated by the LLM.

Field	Content
Hypothesis ID	hyp_w4_d3.0_1
Process Label	senescence_like (Confidence: high)
Summary	“The observed increase in cell area and texture heterogeneity, characteristic of cellular hypertrophy, aligns with a senescence-like state. This is supported by morphological similarity to the perturbation of RRGD, a regulator of mTOR signaling, which is a key driver of senescent cell enlargement and metabolic activity.”
Quant Evidence	obs:obs_w4_d3.0
Retrieval Refs	jump:row_916481, lit:q1:4, lit:q1:6

This example demonstrates identifier-level citation validity and strong proxy compatibility under our evaluation framework. It passes V1 because all cited evidence identifiers are strictly present in the associated payload (record-level validity: `obs_valid_rate=1.0`, `retrieval_valid_rate=1.0`; 0 invalid references). Furthermore, it passes V2 proxy compatibility: under the `senescence_like` mapping (Table A6), 7 top features are proxy hits (`Hit@10 = 0.875`), and these hits account for 86.8% of the absolute delta mass ($WP = 0.868$; Methods, Eq. 5–6). This does not by itself prove that the cited evidence semantically entails the claim, but it shows that the hypothesis is auditable under both V1 and V2. A complete structured example of a global-level claim is provided in Table A7.

4.3 Proxy alignment performance (V2)

We next evaluated whether the LLM-generated `process_label` hypotheses were quantitatively compatible with observed morphology deltas using the predefined proxy mapping (Methods, Eq. 5–6; proxy definitions in Table A6). At the global (dose $\times$ label) level, the mean `Hit@10` over rows was 0.4335 ($n_{\text{dose_label_rows}} = 14$), with 0 missing-delta records.

Dose-stratified results showed a clear dose dependence in best-case compatibility. The per-dose best mean `Hit@10` values were 0.2361 (0.003 mGy/hr), 0.4306 (0.03 mGy/hr), 0.4861 (0.3 mGy/hr), 0.8750 (3.0 mGy/hr), and 0.8889 (6.0 mGy/hr).

Figure 2 visualizes how V2 alignment varies with dose rate for both hierarchical scales (dose-time and global), reporting both the

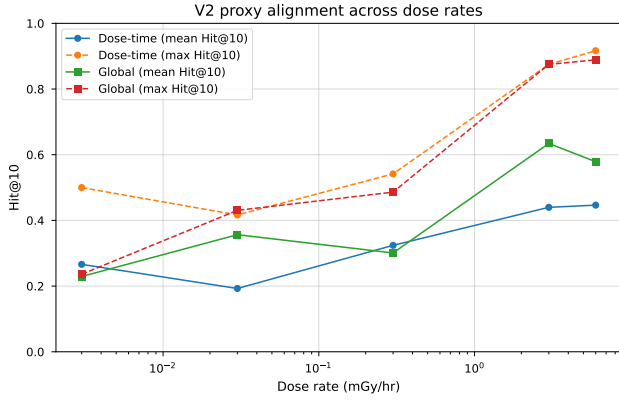


Figure 2: V2 proxy alignment dose-response (Hit@10).

mean Hit@10 across rows and the best-case (max) Hit@10 at each dose. Quantitatively, the global best-case Hit@10 increases from 0.2361 at 0.003 mGy/hr to 0.8889 at 6.0 mGy/hr, while the global mean Hit@10 increases from 0.2292 to 0.5787 over the same range. Dose_time summaries show a similar upward trend (mean Hit@10 from 0.2662 to 0.4464; max Hit@10 from 0.5000 to 0.9167). In the evaluation-first framing, this dose-response matters because it indicates that the compatibility metric is sensitive to perturbation strength: when quantitative deltas are weak, proxy alignment is limited, whereas stronger perturbations permit higher compatibility between declared process_labels and observed morphology signatures.

Table 3: Validation summary (V1 grounding, V2 proxy alignment).

Metric	Value	Notes
V1 grounding integrity		
Week×dose records (successful)	45	Total prompts processed
Hypotheses generated	136	Across all records
Invalid evidence-reference cases	0	V1 citation failures
Mean obs-reference validity	1.000	Valid obs: <id> / total
Mean retrieval-reference validity	1.000	Valid jump: /lit: / total
Coverage (both obs + retrieval refs)	1.000	Fraction of hypotheses
V2 proxy alignment		
Global mean Hit@10 (over rows)	0.4335	Dose×label rows ($n = 14$)
Missing-delta records	0	Should be 0
<i>Per-dose best mean Hit@10 (best row per dose)</i>		
Dose 0.003 mGy/hr	0.2361	
Dose 0.03 mGy/hr	0.4306	
Dose 0.3 mGy/hr	0.4861	
Dose 3.0 mGy/hr	0.8750	
Dose 6.0 mGy/hr	0.8889	

4.4 Hierarchical longitudinal summaries

Beyond week×dose hypotheses, the system produces hierarchical summaries at the dose_time and global levels while preserving

evidence IDs for traceability (Methods). Dose-time summaries partition each dose trajectory into a small set of temporal phases with dominant process labels and supporting evidence.

Figure 3 shows the dose-time integration output as a phase timeline for each dose rate, where contiguous week ranges are grouped into 3–4 phases and assigned dominant process_labels. This visualization is not used as a biological claim; rather, it demonstrates the intended behavior of hierarchical reasoning: the model produces compact, temporally structured summaries that preserve within-dose phase structure instead of collapsing all weeks into a single undifferentiated narrative. Because phases cite week-level observation IDs and retrieval IDs, they remain auditable under V1 and quantifiable under V2.

Figure 4 summarizes the within-week distribution of process_label assignments across all week×dose hypotheses. The non-uniform label usage over time provides a descriptive check that the controlled vocabulary is used in a temporally structured manner (rather than collapsing to a single default label). In the evaluation-first framing, this matters because it motivates why phase-aware aggregation (Figure 3) is preferable to flat averaging: the label distribution can shift across weeks, and hierarchical summaries provide a traceable representation of these shifts.

4.5 Higher-resolution consistency with an LLM-external morphology drift summary

To test whether V2 behaves like a non-arbitrary score, we compared it with the Step 4 morphology drift summary computed without any LLM or retrieval components. We first performed a week×dose analysis by aggregating per-hypothesis V2 scores to the corresponding observation level and comparing them with the matched week×dose drift mean ($n = 45$ observations). For the mean weighted-proxy summary, the association was positive but modest (Pearson $r = 0.307$ with 95% bootstrap CI $[-0.001, 0.566]$; Spearman $\rho = 0.293$ with 95% CI $[-0.021, 0.558]$; bootstrap $n_{boot} = 10000$, seed 42). Mean Hit@10 showed a similar but slightly weaker pattern (Pearson $r = 0.282$; Spearman $\rho = 0.265$).

Figure 5 shows this higher-resolution comparison for the mean weighted-proxy summary. The effect size is intentionally interpreted cautiously: both quantities derive from the same Cell Painting distribution, and the purpose of this analysis is only to test whether V2 co-varies with an LLM-external morphology summary rather than behaving as an arbitrary label-matching score (See Appendix A.1 for a complementary coarse dose-level sanity check).

4.6 Biological interpretation under quantitative constraints.

To provide a more concrete biological interpretation, we explicitly examine how LLM-generated hypotheses vary across dose regimes and how these patterns relate to observed morphological signatures.

At higher dose rates (3.0–6.0 mGy/hr), the dominant hypotheses consistently correspond to senescence-like and DNA damage response (DDR)-related processes. This regime is characterized by strong and coherent morphological shifts, including increases in cell area and perimeter along with pronounced changes in texture

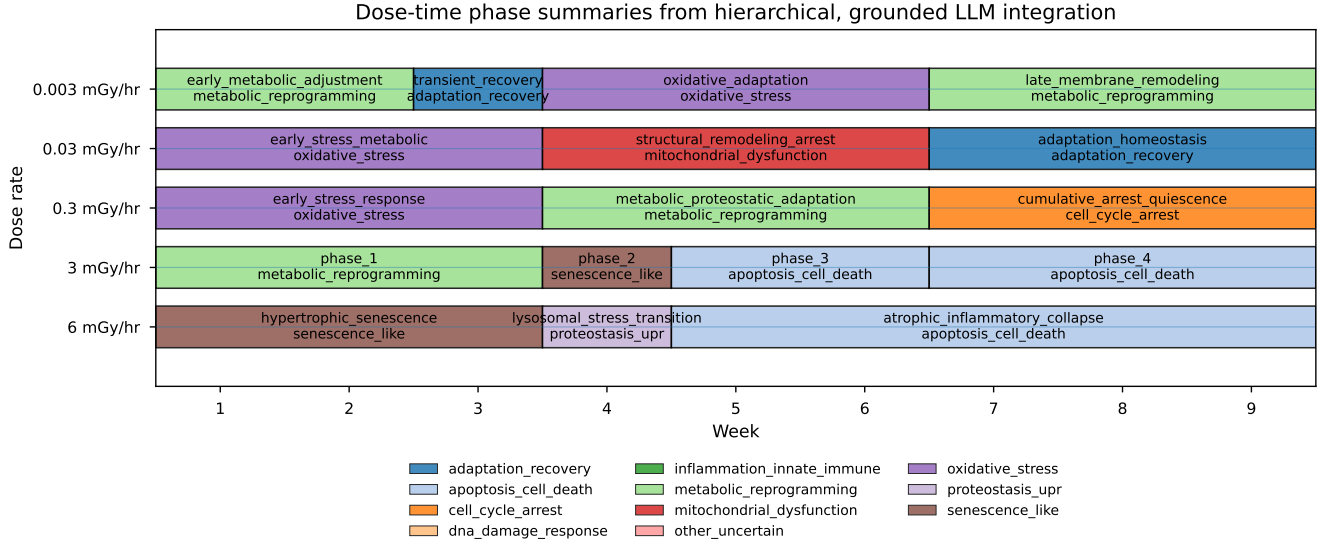


Figure 3: Dose-time integrated phase timeline (hierarchical reasoning output).

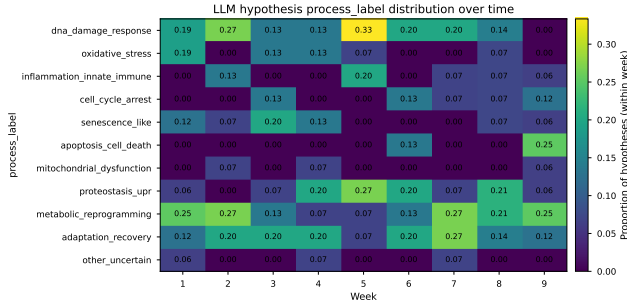


Figure 4: Temporal heatmap of LLM process_label assignments.

features. Such changes are consistent with stress-associated morphological remodeling and hypertrophic phenotypes commonly observed during cellular senescence and growth dysregulation [9, 24].

Importantly, prior work in image-based profiling has demonstrated that morphological features extracted from Cell Painting assays can reliably capture biologically meaningful variation and map to underlying cellular processes [4, 27]. In line with this, we observe substantially higher V2 proxy alignment scores in this high-dose regime, indicating strong quantitative consistency between predicted biological processes and observed morphological changes.

In contrast, at lower dose rates (0.003–0.3 mGy/hr), the model produces more heterogeneous hypotheses, including processes such as metabolic reprogramming, proteostatic stress, and adaptive response programs. In this regime, morphological changes are comparatively subtle and less dominated by canonical stress signatures,

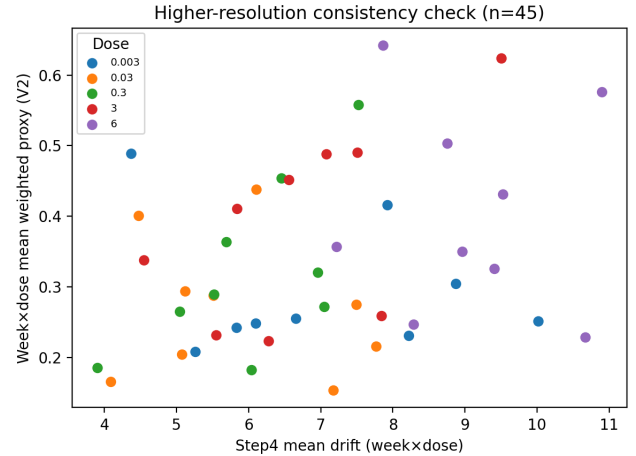


Figure 5: Higher-resolution consistency check between week x dose V2 and the LLM-external morphology drift summary.

leading to lower proxy alignment scores and reduced interpretability of process–morphology relationships.

This pattern is consistent with prior radiobiology literature, which suggests that chronic low-dose radiation often induces non-linear and adaptive cellular responses rather than acute DNA damage-driven phenotypes [22, 31].

5 Discussion

This work is motivated by a practical gap in high-content biology: morphology is information-rich yet difficult to interpret, while

LLMs can generate fluent interpretations that require rigorous auditing before scientific use. We therefore frame LLM outputs as *auditable artifacts*—structured hypotheses with explicit evidence pointers—and emphasize evaluation-first validation via citation validity (V1) and proxy-based morphology compatibility (V2).

Scope and limits of quantitative auditing. V1 enforces *identifier-level* grounding: hypotheses may only cite evidence items that are actually present in the prompt payload. This directly targets out-of-payload references, a known failure mode of neural generation in scientific settings [16, 19, 32]. However, V1 does not guarantee that a cited item semantically supports the claim; it is a necessary but not sufficient condition for scientific faithfulness. V2 complements this by evaluating whether predicted `process_labels` are compatible with the measured morphology deltas under a predefined proxy mapping. This is intentionally conservative: it does not require that the label is *true* in a mechanistic sense, only that it is *quantitatively compatible* with the observed signature. Because the proxy table is hand-specified, V2 should be interpreted as a transparent and reproducible diagnostic rather than as an objective mechanistic benchmark.

Hierarchical reasoning as an auditing strategy. A core design choice is to generate hypotheses at the most atomic unit of measurement (week $\times$ dose) before aggregation. In longitudinal settings, flat aggregation can obscure phase structure and dilute evidence, and it can yield misleading global summaries when trends differ across subgroups (a form of Simpson’s paradox). By preserving evidence IDs from week-level hypotheses into dose_time phases and global summaries, the system maintains traceability while producing compact narratives suitable for downstream review.

Limitations and extensibility. First, V1 evaluates citation validity but not semantic support; a stronger external assessment would require claim–evidence judgments or expert annotation, which we do not yet have. Second, V2 relies on a predefined proxy mapping from `process_label` to expected morphology feature buckets; this mapping is necessarily simplified and may not capture context-dependent morphology–process relationships. Third, the controlled vocabulary constrains expressivity and may collapse distinct biological mechanisms into coarse categories. Fourth, retrieval coverage and bias (e.g., limitations of the reference library, neighbor annotation quality, and pathway text granularity) can shape the evidence available to the LLM and thus the hypotheses produced.

We further emphasize that the use of morphology-only quantitative anchors reflects a practical constraint of the current dataset rather than a limitation of the framework itself. In many longitudinal biological studies—including low-dose radiation experiments—ground-truth mechanism labels and orthogonal modalities are limited or unavailable. Our framework is therefore designed to operate under this constraint, using morphology as a primary quantitative signal while enforcing strict grounding and consistency checks. The auditing structure is inherently extensible: additional modalities such as transcriptomics, proteomics, or targeted biomarkers can be incorporated into the same evidence and validation pipeline to provide stronger mechanistic validation.

Finally, the Step 4 drift summary is derived from the same morphology distribution and should be interpreted only as an LLM-external sanity check, not as independent biological validation.

Implications for low-dose radiation studies. Chronic low-dose and low-dose-rate radiation remains an area where effects can be subtle and context dependent [22, 31]. Our results do not establish mechanisms; rather, they demonstrate that an LLM-centered interpretation workflow can be instrumented with quantitative auditing so that any proposed biological narrative remains tethered to measured morphology and retrievable evidence.

Real-world applicability. Our study is conducted on a real-world longitudinal Cell Painting dataset involving chronic low-dose-rate radiation exposure in RPE-1 cells. This setting reflects a particularly challenging regime in radiobiology, where effects are subtle, context-dependent, and lack well-defined ground-truth labels.

In such scenarios, traditional supervised or benchmarking approaches are not feasible. Our framework is specifically designed for this setting: it leverages limited experimental data and external knowledge to generate structured, auditable hypotheses that can guide downstream experimental validation.

This makes the approach directly applicable to other real-world biological contexts characterized by limited labels and complex temporal dynamics, such as drug response profiling, aging studies, and environmental exposure analysis.

6 Conclusion

We presented an LLM-centered, retrieval-augmented interpretation framework for longitudinal Cell Painting morphology data that is evaluation-first by design. By enforcing structured JSON outputs with controlled labels and explicit evidence references, and by validating outputs using grounding integrity (V1) and proxy-based process–morphology alignment (V2), the framework enables traceable hypothesis generation that can be quantitatively audited against independent morphology signals.

More broadly, this work contributes a practical template for integrating LLMs into computational biology workflows where narrative interpretation must remain accountable to measured data: hypotheses are generated with explicit provenance, assessed quantitatively, and summarized hierarchically without sacrificing traceability.

7 Acknowledgment

This material is based upon works for the LUCID: Low-dose Understanding, Cellular Insights, and Molecular Discoveries program under Award Number DE-AC02-06CH11357 supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research.

References

- [1] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.
- [2] Daniil A Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332* (2023).
- [3] Mark-Anthony Bray, Shantanu Singh, Han Han, Chad T. Davis, Blake Borgeson, Carol Hartland, Maria Kost-Alimova, Sigrun M. Gustafsdottir, Christopher C. Gibson, and Anne E. Carpenter. 2016. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols* 11, 9 (2016), 1757–1774. doi:10.1038/nprot.2016.105

- [4] Juan C Caicedo, Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S Vasilevich, Joseph D Barry, Harmanjit Singh Bansal, Oren Kraus, et al. 2017. Data-analysis strategies for image-based cell profiling. *Nature methods* 14, 9 (2017), 849–863.
- [5] Srinivas Niranj Chandrasekaran, Beth A. Cimini, Amy Goodale, Lisa Miller, Maria Kost-Alimova, Nasim Jamali, John G. Doench, Briana Fritchman, Adam Skepner, Michelle Melanson, Alexandr A. Kalinin, John Arevalo, Marzieh Haghighi, Juan C. Caicedo, Daniel Kuhn, Desiree Hernandez, James Berstler, Hamdah Shafqat-Abbasi, David E. Root, Susanne E. Swalley, Sakshi Garg, Shantanu Singh, and Anne E. Carpenter. 2024. Three million images and morphological profiles of cells treated with matched chemical and genetic perturbations. *Nature Methods* 21 (2024), 1114–1121. doi:10.1038/s41592-024-02241-6 Published online 2024-04-09.
- [6] Europe PMC Consortium. 2015. Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic acids research* 43, D1 (2015), D1042–D1048.
- [7] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92. doi:10.1145/3458723 Originally released as arXiv:1803.09010.
- [8] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. PubMedBERT: Domain-Specific Language Model Pretraining for Biomedical Text Mining. arXiv:2007.15779 [cs.CL] doi:10.48550/arXiv.2007.15779 TO BE VERIFIED: if citing a published version, replace arXiv with venue details.
- [9] Alejandra Hernandez-Segura, Jamil Nehme, and Marco Demaria. 2018. Hallmarks of cellular senescence. *Trends in cell biology* 28, 6 (2018), 436–453.
- [10] Arturo Hernandez-Segura, Justine Nehme, and Marco Demaria. 2018. New hallmarks of cellular senescence. *Trends in Cell Biology* 28, 6 (2018), 436–453.
- [11] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. arXiv:2311.05232 [cs.CL] doi:10.48550/arXiv.2311.05232 arXiv:2311.05232 is a hallucination survey (not specifically a fabricated-citations-only study)..
- [12] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE transactions on big data* 7, 3 (2019), 535–547.
- [13] Guido Kroemer, Lorenzo Galluzzi, Peter Vandenabeele, John Abrams, Emad S Alnemri, Eric H Baehrecke, Mikhail V Blagosklonny, Wafik S El-Deiry, Pierre Golstein, Douglas R Green, et al. 2009. Classification of cell death: recommendations of the Nomenclature Committee on Cell Death 2009. *Cell Death & Differentiation* 16, 1 (2009), 3–11.
- [14] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chanho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240. doi:10.1093/bioinformatics/btz682
- [15] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html Proceedings version (NeurIPS 2020). arXiv:2005.11401 is the preprint..
- [16] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*. 3214–3252.
- [17] Renqian Luo, Litao Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. arXiv:2210.10341 [cs.CL] doi:10.48550/arXiv.2210.10341
- [18] Markus Marks, Uriah Israel, Rohit Dilip, Qilin Li, Changhua Yu, Emily Laubscher, Ahamed Iqbal, Elora Pradhan, Ada Ates, Martin Abt, et al. 2025. CellSAM: a foundation model for cell segmentation. *Nature Methods* (2025), 1–9.
- [19] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 1906–1919.
- [20] Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, Robert Petryszak, Eliot Ragueneau, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Thawfeek Varusai, Joel Weiser, Adam Wright, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. 2024. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research* 52, D1 (2024), D672–D678. doi:10.1093/nar/gkad1025
- [21] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. 220–229. doi:10.1145/3287560.3287596
- [22] National Research Council. 2006. *Health Risks from Exposure to Low Levels of Ionizing Radiation: BEIR VII Phase 2*. The National Academies Press, Washington, DC. doi:10.17226/11340
- [23] Gabriel E Neurohr, Rachel L Terry, Jette Lengefeld, Megan Bonney, Gregory P Brittingham, Fabrice Moretto, Teemu P Miettinen, Laura P Vaites, Lidia M Soares, Joao A Paulo, et al. 2019. Excessive cell growth causes cytoplasm dilution and contributes to senescence. *Cell* 176, 5 (2019), 1083–1097.
- [24] Gabriel E Neurohr, Rachel L Terry, Jette Lengefeld, Megan Bonney, Gregory P Brittingham, Fabien Moretto, Teemu P Miettinen, Laura Pontano Vaites, Luis M Soares, Joao A Paulo, et al. 2019. Excessive cell growth causes cytoplasm dilution and contributes to senescence. *cell* 176, 5 (2019), 1083–1097.
- [25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.
- [26] Mohammad Hasan Rohban, Shantanu Singh, Xiaoyun Wu, Julia B Berthet, Mark-Anthony Bray, Matthew D Shair, Lee L Rubin, and Anne E Carpenter. 2017. Systematic morphological profiling of human gene and allele function via Cell Painting. *eLife* 6 (2017), e24060.
- [27] Mohammad Hossein Rohban, Shantanu Singh, Xiaoyun Wu, Julia B Berthet, Mark-Anthony Bray, Yashaswi Shrestha, Xaralabos Varelas, Jesse S Boehm, and Anne E Carpenter. 2017. Systematic morphological profiling of human gene and allele function via Cell Painting. *Elife* 6 (2017), e24060.
- [28] Maria Janina Sarol, Shufan Ming, Shruthan Radhakrishna, Jodi Schneider, and Halil Kilicoglu. 2024. Assessing citation integrity in biomedical publications: corpus annotation and NLP models. *Bioinformatics* 40, 7 (2024), btae420. doi:10.1093/bioinformatics/btae420 Published 2024-06-26.
- [29] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, et al. 2023. Large language models encode clinical knowledge. *Nature* 620, 7972 (2023), 172–180.
- [30] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
- [31] United Nations Scientific Committee on the Effects of Atomic Radiation. 2022. *Sources, Effects and Risks of Ionizing Radiation: UNSCEAR 2020/2021 Report to the General Assembly, with Scientific Annexes, Volume III – Scientific Annex C: Biological mechanisms relevant for the inference of cancer risks from low-dose and low-dose-rate radiation*. Technical Report. United Nations. Scientific Annex C (low-dose/low-dose-rate). A corrigendum was issued in Sep 2022..
- [32] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or Fiction: Verifying Scientific Claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7534–7550. doi:10.18653/v1/2020.emnlp-main.609

A Supplementary Materials

Table A1: Dataset scale and preprocessing summary.

Item	Value	Notes
Biological Setup & Experimental Design		
Cell line	RPE-1	Human retinal pigment epithelial
Assay	Cell Painting	High-content morphology
Plate format	96-well	
Time course	Weeks 1–9	9 longitudinal time points
Dose rates (mGy/hr)	0.003, 0.03, 0.3, 3.0, 6.0	5 treated dose conditions
Paired controls	Plate C (weekly)	Paired control plate per week
Treated replicates	4 wells / (week, dose)	Fixed plate map
Total treated wells	180	9 weeks $\times$ 5 doses $\times$ 4 wells
Total control wells	34	Week 1: 2; Weeks 2–9: 4
Total wells analyzed	214	Treated + control well profiles
Data Dimensions		
Fields of view (FOV) / well	9	$f \in \{1, \dots, 9\}$
Z-planes / FOV	15	$p \in \{1, \dots, 15\}$
Nominal CSV files / well	135	9×15 (when complete)
Total cell-level CSV files	28,886	(4 files excluded via QC)
Total cell-level observations	6,375,699	Sum of per-well cell counts
Features & Preprocessing		
Excluded columns	orientation, centroid_x, centroid_y	Excluded at well-level aggregation
Base morphology feature set	9 features	area, perimeter, eccentricity, solidity, mean_intensity, and 4 glcm texture features
Well-level summary features	36	9 base $\times$ {median, MAD, q10, q90}
Within-week normalization	Robust Z-score	$(x - \text{median}) / (\text{MAD} + 10^{-6})$
Delta design	Treated – Control (weekly paired)	$\Delta Z = Z - \text{median}(Z_{\text{control}}, \text{week})$

Table A2: System configuration summary (retrieval and LLM settings).

Component	Parameter	Value
Retrieval (JUMP)	Reduced embedding	9 morphology features; z-score standardization ($\epsilon = 10^{-8}$)
	Neighbor search backend	FAISS IndexFlatL2 (if available); else sklearn NearestNeighbors (Euclidean)
	Candidate neighbors (k)	100
	Final neighbors retained	10
	Min. gene-like neighbors	3
Retrieval (Reactome)	Gene $\rightarrow$ pathway top N	12
	Pathway document search	TF-IDF (1–2 grams); cosine similarity
	TF-IDF	250,000
	max_features	
	TF-IDF min_df	2
	Document retrieval k	6
	Reactome release	Downloaded from /current/ endpoint
LLM Reasoner	Model	gemini-3-pro-preview
	Temperature	0.2
	Max output tokens	8192
	Thinking level	HIGH
	Output format	JSON-only; schema-enforced
	Controlled vocabulary	11 canonical process_labels
Grounding	Strict grounding	Enabled (invalid/missing references $\rightarrow$ reject & retry)
	Citation types (week $\times$ dose)	obs:<id>, jump:<id>, lit:<id>
	Citation types (hierarchical)	obs:, jump:, lit:, path:<Reactome_id>
Hierarchy	week $\times$ dose $\rightarrow$ dose_time	Evidence preserved with per-week catalogs
	dose_time $\rightarrow$ global	Cross-dose summarization with evidence-ID propagation

Table A3: Delta and baseline diagnostics.

Feature (Δ z-score)	Control		Treated	
	Median $ \Delta $	q95 $ \Delta $	Median $ \Delta $	q95 $ \Delta $
area_median	0.520	2.738	2.080	6.484
perimeter_median	0.374	2.550	2.322	6.008
eccentricity_median	0.545	1.517	1.379	4.745
solidity_median	1.041	2.383	1.399	4.758
mean_intensity_median	0.533	1.473	1.229	3.892
glcm_contrast_median	0.728	2.503	1.260	4.500
glcm_correlation_median	0.899	3.124	1.489	5.199
glcm_energy_median	0.425	2.736	2.164	5.933
glcm_homogeneity_median	0.433	1.570	1.871	4.544

Note: Computed on well-level delta profiles.

($\Delta = z - \text{median}(z_{\text{control}}, \text{week})$).

$n_{\text{control}} = 34$, $n_{\text{treated}} = 180$.

Table A4: Process label usage summary.

Process label	week×dose hypotheses	dose_time phases	Global claims	Global dose summaries
dna_damage_response	24 (17.6%)	1 (5.9%)	1 (25.0%)	2 (40.0%)
oxidative_stress	9 (6.6%)	4 (23.5%)	1 (25.0%)	0 (0.0%)
inflammation_innate_immune	8 (5.9%)	3 (17.6%)	1 (25.0%)	1 (20.0%)
cell_cycle_arrest	8 (5.9%)	6 (35.3%)	0 (0.0%)	2 (40.0%)
senescence_like	10 (7.4%)	2 (11.8%)	1 (25.0%)	2 (40.0%)
apoptosis_cell_death	6 (4.4%)	4 (23.5%)	1 (25.0%)	2 (40.0%)
mitochondrial_dysfunction	3 (2.2%)	3 (17.6%)	1 (25.0%)	0 (0.0%)
proteostasis_upr	17 (12.5%)	5 (29.4%)	1 (25.0%)	2 (40.0%)
metabolic_reprogramming	25 (18.4%)	8 (47.1%)	1 (25.0%)	2 (40.0%)
adaptation_recovery	23 (16.9%)	4 (23.5%)	1 (25.0%)	1 (20.0%)
other_uncertain	3 (2.2%)	2 (11.8%)	0 (0.0%)	0 (0.0%)

Note: Denominators: week×dose hypotheses $n = 136$; dose_time phases $n = 17$; global claims $n = 4$; global dose summaries $n = 5$.

Table A5: Grounding reference statistics (evidence-reference coverage and counts).

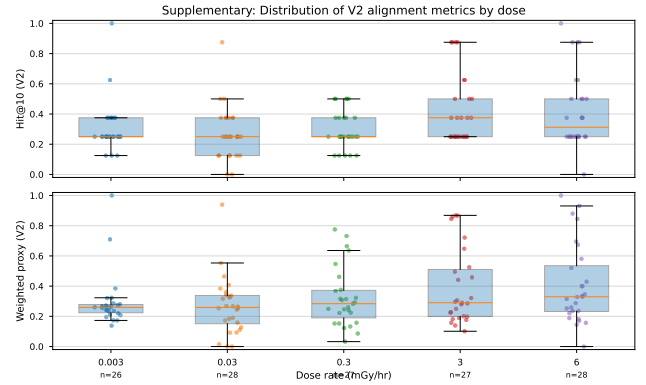
Metric	Value	Notes
Run scale		
Week×dose records	45	Prompt instances
Hypotheses (total)	136	Across all records
Hypotheses per record (mean)	3.022	
Hypotheses per record (min–max)	2–4	Distribution: {2:1, 3:42, 4:2}
Evidence references per hypothesis		
Quant refs per hypothesis (mean)	1.000	obs:<id> citations
Quant refs per hypothesis (min–max)	1–1	
Retrieval refs per hypothesis (mean)	2.838	jump:/lit: citations
Retrieval refs per hypothesis (min–max)	1–5	
Retrieval evidence usage		
Hypotheses citing JUMP neighbors	124 (91.2%)	At least one jump: ref
Hypotheses citing literature	102 (75.0%)	At least one lit: ref
Hypotheses citing both jump + lit	90 (66.2%)	
Total retrieval refs (jump)	219	Count across all hypotheses
Total retrieval refs (lit)	167	Count across all hypotheses
Grounding caps (per record)		
Allowed obs IDs	1	
Allowed JUMP refs	10	
Allowed literature refs	10	

Table A6: V2 proxy mapping (process labels to proxy feature sets).

Process label	Proxy expectations (reduced morphology feature set)
dna_damage_response	glcm_contrast(+), glcm_correlation(0), mean_intensity(0), area(0)
oxidative_stress	glcm_contrast(+), mean_intensity(0), area(0)
inflammation_innate_immune	glcm_contrast(+), area(0)
cell_cycle_arrest	area(+), perimeter(+), eccentricity(0), solidity(0)
senescence_like	area(+), perimeter(+), solidity(0), eccentricity(0), glcm_contrast(+), glcm_correlation(0), glcm_energy(-), mean_intensity(0)
apoptosis_cell_death	area(-), perimeter(-), mean_intensity(0), glcm_contrast(+), glcm_energy(-)
mitochondrial_dysfunction	mean_intensity(0), glcm_contrast(+)
proteostasis_upr	glcm_contrast(+), glcm_energy(-)
metabolic_reprogramming	mean_intensity(0), area(0)
adaptation_recovery	area(0), glcm_contrast(0)
other_uncertain	(none)

Note: Direction codes: (+) expected increase; (-) expected decrease; (0) either direction / weak proxy. Feature mappings are derived from canonical morphological hallmarks of cell stress [10, 13, 26]. Used strictly as a quantitative consistency check (V2), not as a biological ground truth.

Used as a consistency check, not biological ground truth.

**Figure A1: V2 metric distributions.**

A.1 Coarse Dose-Level Sanity Check

As a complementary coarse summary to the higher-resolution week×dose analysis ($n = 45$) presented in the main text (Section 4.5), we also evaluated a dose-level comparison ($n = 5$ doses). As shown

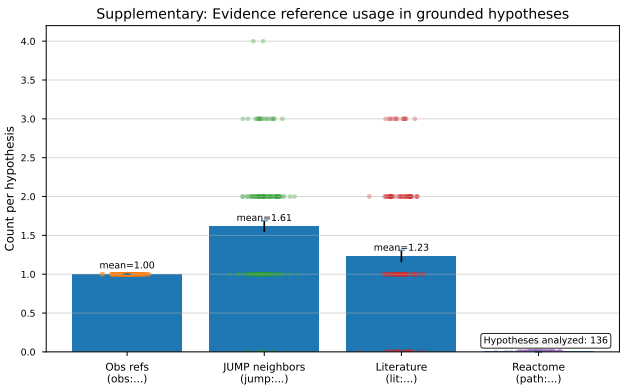


Figure A2: Evidence reference composition.

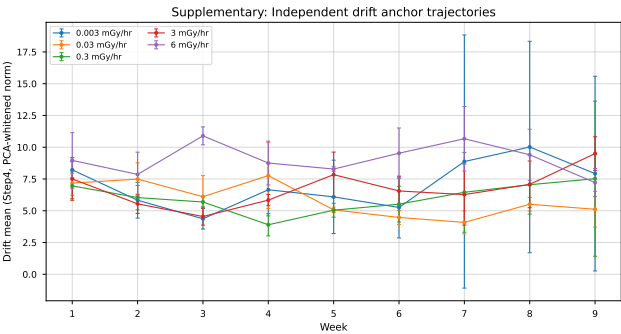


Figure A3: Step 4 drift trajectories.

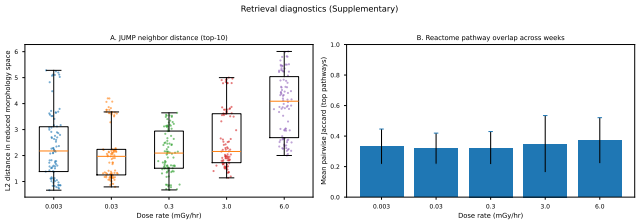


Figure A4: Retrieval diagnostics.

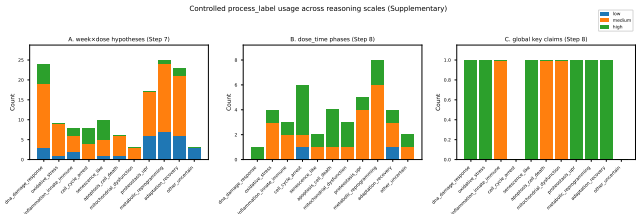


Figure A5: Controlled vocabulary usage across reasoning scales.

Table A7: Structured Example of a Global-Level Integrated Claim

Field	Content
Claim ID	claim_low_dose_adaptation
Claim Statement	Chronic low-dose radiation (0.003–0.3 mGy/hr) induces a distinct adaptive phenotype characterized by metabolic reprogramming and proteostatic stress, contrasting with the hypertrophic senescence seen at higher doses.
Process Labels	metabolic_reprogramming, adaptation_recovery, proteostasis_upr
Quantitative Evidence	obs:obs_w6_d0.003, obs:obs_w9_d0.03, obs:obs_w7_d0.3
Retrieval References	jump:row_862582, jump:row_928254, lit:q1:1
Confidence Level	High

in Figure A6, the association remained positive for the dose_time mean Hit@10 summary (Pearson $r = 0.661$ with 95% bootstrap CI $[-0.363, 1.000]$; Spearman $\rho = 0.700$ with 95% CI $[-0.875, 1.000]$).

Given the extremely small sample size, the confidence intervals are understandably wide. We retain this five-point analysis strictly as a low-resolution sanity check. Consistent with our approach in the main text, we do not interpret this as an independent biological validation, but rather as a basic verification that the model's scores co-vary positively with the expected morphological drift at the macro-dose level.

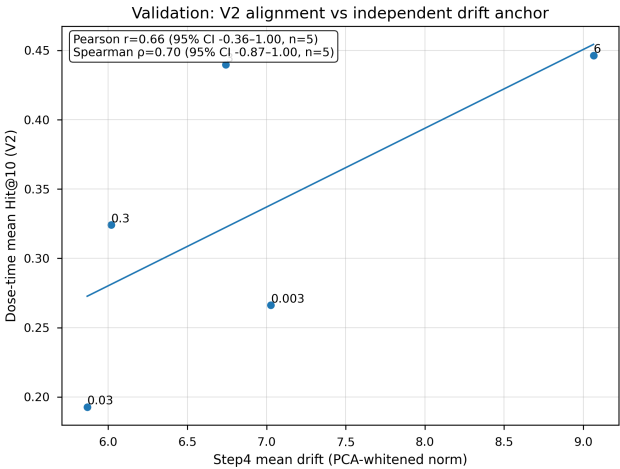


Figure A6: Coarse dose-level comparison ($n = 5$) evaluating the consistency between the LLM-external morphology drift and the dose_time mean Hit@10 summary. While a positive trend is visible, the small sample size limits statistical confidence.