A Knowledge-grounded Agentic Framework for Single-cell RNA Sequencing Method Hypothesis Generation

Xuewen Zhou[1] **Haiman Guo**[1] **Lichao Sun**[2] **Zhen Chen**[1,*]
[1] The Hong Kong Polytechnic University [2] Lehigh University

This Supplementary Information provides supporting protocols, numerical summaries and worked comparisons for the analyses reported in the main text.

Supplementary analyses of RPKG-derived hypothesis formulation

Blinded human selection of candidate hypotheses

For each of 30 Problems, we prepared a triplet of candidate method hypotheses from three generation conditions, following the blinded hypothesis-evaluation protocol in the main-text Methods. The full scHypothesizer condition used an RPKG-derived Problem–Mechanism pairing, its linked literature evidence and the complete Proposer-Critic process. The Abstract-only condition received the same Problem together with the original paper abstracts. The Problem-only condition received only the Problem description. The three candidates were matched in length and presentation style, stripped of condition-revealing labels and randomly ordered.

Six reviewers, comprising three doctoral-level and three master's-level researchers, independently selected one candidate from every triplet. The instruction asked which candidate they would be most willing to pursue for further method development. Reviewers had to choose exactly one candidate; ties, abstention, and a "none of these" response were not permitted. The design yielded 180 decisions, whose anonymized decision-level records are provided as Supplementary Data 3.

Table S1 | Blinded human selection of candidate hypotheses. Each generation condition appeared once in every one of the 30 candidate triplets evaluated by six reviewers. Counts are forced-choice decisions.

Topic	Problems	scHypothesizer	Abstract-only	Problem-only
Ambient	10	37	12	11
Integration	10	33	17	10
Reference mapping	10	28	16	16
All topics	30	98 (54.4%)	45 (25.0%)	37 (20.6%)

Supplementary Note 1: MCG organizes retrieved evidence around the matched Mechanism

Purpose and scope. This analysis examined whether MCG consistently organized retrieved evidence around the matched Mechanism when compared with simple RAG using the same literature corpus. It characterized how proposals were structured; a multi-component design is not in itself inappropriate.

Matched comparison. Following the matched-construction protocol in the main-text Methods, simple RAG and pre-critic MCG used the same language-model backend, output schema, literature corpus, complete novelty summaries and single generation pass for each of the same 30 Problems. Simple RAG retrieved the five papers whose titles were most similar to the Problem and conditioned the generator on their novelty summaries. Pre-critic MCG instead selected summaries through the typed graph pairing and used the matched Mechanism as the candidate's organizing causal operation. The pre-critic condition omitted the subsequent Proposer-Critic loop, thereby separating evidence selection and initial construction from subsequent refinement.

Descriptive coding. The paired candidates were examined side by side with method names masked. A candidate was coded as *module concatenation* when it chained components from several retrieved methods without one causal operation organizing the proposal. A candidate was coded as *organized around the matched Mechanism* when the graph-selected Mechanism remained its central causal operation. Auxiliary implementation steps did not change this designation when they supported that operation.

Table S2 | Structure of candidates in the matched construction comparison. Counts report Problems for which simple RAG was coded as module concatenation and pre-critic MCG was organized around the matched Mechanism.

Topic	Problems	Simple RAG: module concatenation	Pre-critic MCG: matched Mechanism
Ambient	10	9	10
Integration	10	7	10
Reference mapping	10	7	10
All topics	30	23	30

Worked comparisons

The following worked cases illustrate the coding. Each shows the simple RAG candidate, the pre-critic MCG candidate, and the basis for the structural distinction, with method names masked.

Case 1: Ambient, sc-DGE-insufficient correction. **Simple RAG.** The candidate estimated sample- and gene-specific contamination and applied correction only where contamination was frequent enough to affect testing. It formed an ambient profile from low-UMI droplets; when these were absent, it flagged candidate ambient genes using concentrated low-count patterns within preclustered populations. It then set a conservative subtraction threshold, gated eligibility by an ambient-prevalence proxy and added an optional pilot differential-expression screen.

Pre-critic MCG. The candidate used one per-cell mixture of native and contaminating multinomial expression. The contaminant was a weighted combination of other populations' profiles in the same sample, excluding the cell's own population. A single variational expectation-maximization procedure estimated native profiles, contamination fractions and source-population weights.

Structural contrast. The simple RAG candidate passes data through four separate stages (detector, threshold rule, eligibility gate, and screen) that share no common estimated quantity, whereas every step of the pre-critic MCG candidate depends on the per-cell contamination mixture.

Case 2: Ambient, experiment-specific contamination profiles. **Simple RAG.** The candidate combined low-count droplet libraries with expression irregularities within clustered populations. It derived a sample ambient profile, identified candidate genes using widespread low-entropy counts, estimated thresholds from putatively non-expressing groups and applied adaptive subtraction. An additional empirical-shrinkage stage pooled information across technically similar experiments.

Pre-critic MCG. The candidate used a hierarchical count mixture fitted to raw UMI counts. A per-sample Dirichlet-multinomial ambient profile, per-cell contamination fraction and gene-level native-versus-ambient posterior jointly determined corrected counts.

Structural contrast. In the simple RAG candidate, two evidence sources feed successive correction stages; in the pre-critic MCG candidate, the corrected matrix is read directly from the posterior of the hierarchical mixture.

Case 3: Ambient, uncertain contamination-rate estimation. **Simple RAG.** The candidate combined ambient droplet profiles, internal low-expression reference groups, nuclear-integrity features, allele-informative counts and gene-wise entropy deviations. A hierarchical model combined these estimators into posterior intervals and exposed a conservative-to-aggressive correction path scored by several diagnostics.

Pre-critic MCG. The candidate used a Bayesian mixture in which each count was native or contaminating. Each cell carried a posterior over its contamination fraction, and a user-selected posterior quantile controlled correction aggressiveness.

Structural contrast. Simple RAG sought robustness by combining heterogeneous estimators. Pre-critic MCG represented uncertainty through one mixture posterior and its quantiles.

Case 4: Ambient, mixed-severity contamination. **Simple RAG.** The candidate proposed two correction layers. One scored gene severity and subtracted gene-specific thresholds from high-severity genes. A second fitted a residual native-versus-contaminating mixture to remove diffuse background. Because both layers acted on the same counts, a third reconciler was added to prevent double subtraction.

Pre-critic MCG. The candidate used one mixture of endogenous transcription and a sample-specific ambient profile. Correction was gene-adaptive within this model, with targeted subtraction for strongly ambient genes and capped subtraction for weak diffuse signals.

Structural contrast. Because its two correction layers act on the same counts, the simple RAG candidate needs a third module to prevent double subtraction; the pre-critic MCG candidate treats severe and diffuse contamination as two regimes of the same mixture.

Case 5: Ambient, a second uncertain contamination-rate problem. **Simple RAG.** The candidate estimated contamination intervals at sample, cell and gene levels using empty-droplet profiles, filtered-cell reference groups, droplet-quality features and gene-distribution evidence. It then exposed conservative, nominal and aggressive correction settings with marker-preservation constraints.

Pre-critic MCG. The candidate used a multinomial mixture of endogenous expression and a batch-specific ambient profile. A per-cell contamination posterior and user-controlled quantile defined the correction strength while protecting genes with strong endogenous support.

Structural contrast. Simple RAG assembled several estimators to express uncertainty. Pre-critic MCG expressed uncertainty through one mixture posterior and a quantile.

Case 6: Ambient, contamination in cardiac snRNA-seq. **Simple RAG.** The candidate jointly performed nuclei calling and ambient correction. It calculated several droplet-quality features, clustered droplets, removed reference-cluster outliers, fitted a native-plus-ambient mixture and flagged gene-specific contamination using count and entropy patterns.

Pre-critic MCG. The candidate used per-sample adaptive ambient profiling. It swept low-UMI thresholds, selected the threshold that best explained ectopic cardiac-marker expression in unexpected populations and then applied capped gene-wise subtraction.

Structural contrast. The simple RAG candidate runs calling, clustering, outlier removal, and two correction stages in sequence; in the pre-critic MCG candidate, all downstream steps follow from the choice of the empty-library threshold.

Case 7: Integration, alignment under limited cell overlap. **Simple RAG.** The candidate denoised each batch into a low-rank space, tested whether a shared subspace was plausible, created provisional anchors, expanded and refined them, learned a shape-preserving transformation and assigned an unaligned status from residuals.

Pre-critic MCG. The candidate used a count variational autoencoder with batch-balanced mini-batches and encoder Batch Normalization. The normalization statistics estimated a pooled cell-state distribution, creating a common coordinate system without explicit anchor matching.

Structural contrast. Alignment in the simple RAG candidate emerges from a sequence of estimation, anchoring, transformation, and filtering steps; in the pre-critic MCG candidate it depends on how normalization statistics are computed inside the encoder.

Case 8: Integration, batch and technology effects. **Simple RAG.** The candidate combined an adversarial batch discriminator, a local maximum-mean-discrepancy penalty and a counterfactual cycle-consistency loss with reconstruction, a mixture prior and separately refined anchors.

Pre-critic MCG. The candidate used a batch- and technology-conditioned negative-binomial autoencoder. Batch effects entered the decoder, while reconstruction fidelity remained the primary organizing constraint; strong adversarial and distribution-matching terms were omitted.

Structural contrast. The simple RAG candidate must balance three invariance objectives; the pre-critic MCG candidate relies on conditioning the decoder, and not the encoder, on technical labels.

Case 9: Integration, intercellular interactions ignored by independent-cell modeling. **Simple RAG.** The candidate represented each sample as a graph and combined masked graph attention, shared and private encoders, adversarial alignment, orthogonality penalties, graph decoders, supervised loss and entropy minimization.

Pre-critic MCG. The candidate used one graph-and-feature autoencoder that reconstructed expression with a count likelihood and adjacency with an edge decoder. Batch invariance and additional weighting for rare or topologically distinctive neighborhoods supported the same joint-reconstruction mechanism.

Structural contrast. The simple RAG candidate stacks encoders, alignment terms, and decoders; the pre-critic MCG candidate is built on joint reconstruction of expression and graph structure.

Case 10: Reference mapping, labor-intensive manual cell labeling. **Simple RAG.** The candidate combined several reference classifiers, a denoising count autoencoder, prediction agreement, entropy, reconstruction anomaly and neighborhood consistency. These signals were calibrated and passed to a mixture model that separated confident known cells from unresolved cells.

Pre-critic MCG. The candidate partitioned genes into stable identity genes, essential markers and nuisance genes before fitting a classifier. Query cells were compared with type prototypes, and an essential-marker gate supported explicit abstention.

Structural contrast. The simple RAG candidate fuses several classifiers and uncertainty signals; the pre-critic MCG candidate rests on partitioning genes by evidence type, from which the marker gate and abstention follow.

Supplementary Note 2: Empirical assessment evidence profiles

The four researcher-selected candidates entering Empirical Assessment were evaluated with benchmark plans fixed before execution, following the executable-case scoring protocol in the main-text Methods. Each plan separated performance on the conventional task from the capability proposed by the method hypothesis. The resulting board scores are case-specific composites. Within each dataset–task–metric combination with values for at least two methods, metrics were oriented so that higher values indicated better performance and were min–max scaled before aggregation. Combinations in which a metric was not applicable were excluded, and no missing values were imputed; a method whose own implementation failed on an eligible combination received zero. The composites support ranking within a case and are not comparable across cases, so the underlying metrics are reported alongside each composite.

Table S3 | Summary of the four empirical evidence profiles. The final column summarizes what the assessment supported.

Candidate	Targeted evidence	General board	Supported interpretation
AmbientCap	0.97; PBMC rare-cell kNN overlap 0.4353 versus 0.1238	0.67, rank 2	Rare-cell preservation was supported in the evaluated droplet settings; marker retention weakened in the stated single-nucleus and broad-label boundary settings.
scOSCAR	Novel-cell detection 0.73, rank 5 of 10; coverage-under-shift composite 0.74 versus 0.26 for CSA	0.82, rank 4	Calibrated coverage under shift was supported; novel-cell detection did not exceed that of established mappers and was weakest for near-lineage novelty.
CellWDA	0.78; shared-support AUROC 0.836 versus 0.809	0.64, rank 1	Selective alignment was supported in the evaluated partial-overlap settings, with weaker recovery in the donor-only boundary setting.
CONFIDE-VI	Central alignability AUROC 0.47 versus chance 0.50; secondary scaffold composite 0.83	0.57, rank 4	The central alignability claim was not supported; secondary density-stratified measurements showed limited signal.

Per-method Empirical Assessment data sheets

The following four data sheets summarize the evidence for the researcher-selected candidates. Each records the RPKG-derived Problem, the proposed Mechanism and its literature grounding, the general and targeted evaluation settings, datasets, comparators, observed results, and the settings in which the claim was not supported.

AmbientCap — Ambient RNA decontamination

Topic: Rare-cell-preserving ambient RNA decontamination

Problem (from the RPKG).
A transcript can be genuine ambient contamination in the dominant population yet a true marker in a rare one, so uniform correction can erase real rare-cell signal.

Proposed Mechanism.
Cluster-conditional ambient-null projection with a per-(gene, cluster) protection cap. *Grounded in: DGE-aware ambient gene filtering.*

General performance.
The overall decontamination board measured standard removal quality on designed mixtures and reference cohorts. Metrics were contamination-fraction RMSE, Kendall rank correlation, ectopic-marker detection, differential-expression false positives, marker retention, cell-type silhouette and over-correction rate.

Targeted capability.
Rare-cell preservation under ambient challenge. Metrics were rare-cell k-nearest-neighbour overlap and cell-type silhouette.

Datasets.

- **Mouse kidney strain mixture**: mouse, scRNA-seq; GEO GSE218853; general.
- **Human–mouse barnyard (HEK293T/NIH3T3)**: human and mouse, 10x Genomics; Zenodo 10.5281/zenodo.13831774; general.
- **PBMC 10k (healthy donor)**: human, 10x Genomics; Zenodo 10.5281/zenodo.19581816; targeted.
- **LungMAP broad-age lung**: human, snRNA-seq; CZ CELLxGENE collection 3de0ad6d-4378-4f62-b37b-ec0b75a50d94; targeted.
- **Lukassen lung and airway**: human, scRNA-seq; CZ CELLxGENE collection 2f132ec9-24b5-422f-9be0-ccef03b4fe28; targeted.
- **Mouse mammary ageing atlas**: mouse, scRNA-seq; GEO GSE216542; targeted.

Comparators (method: family).

- SoupX[1], DecontX[2], scCDC[3] and FastCAR[4]: dedicated ambient-RNA decontamination methods using profile- or Bayesian-contamination estimation.

Observed evidence.

- **Targeted capability, rare-cell preservation.** AmbientCap ranked first on the targeted board (0.97). It ranked first for k-nearest-neighbour overlap in all four cohorts (0.4353 versus 0.1238 for the strongest comparator on PBMC) and first for cell-type silhouette in all four cohorts.
- **General performance, ambient removal.** It ranked second on the overall decontamination board (0.67) and obtained the lowest kidney contamination RMSE (0.1207 versus 0.1340 for the next method).
- **General performance, rank ordering.** Kendall correlation was slightly below DecontX and below scCDC on the kidney mixture.
- **Operating boundary.** Marker-magnitude retention was lower in the single-nucleus and broad-label settings (0.9235 on LungMAP versus approximately 0.99 for the strongest comparator), so these settings were not supported as a default use case.

scOSCAR Open-set reference mapping with calibrated uncertainty

Topic: Reference mapping with calibrated uncertainty under donor and study shifts

Problem (from the RPKG).
Closed-set reference mappers assign confident labels to query cells that lie near a known lineage but represent an unseen state, or to known states displaced by donor, compartment or study effects.

Proposed Mechanism.
Candidate-conditioned conformal p-values, a donor-aware Wasserstein envelope and a support-first set-valued decision that rejects a cell when no reference class is supported. *Grounded in: conformal out-of-distribution control.*

General performance.
Closed-set reference-mapping accuracy on cohorts in which every query cell had a matching class in the reference. The metric was known-cell macro-F1.

Targeted capability.
The targeted capability had two parts. The first was novel-cell detection: held-out cell types were removed from the reference, and the method had to flag them as novel while retaining labels for known cells. The board combined six metrics: AUROC, AUPRC, hard-novel recall, known-cell false-rejection rate, per-donor false-rejection rate, and accepted-known macro-F1, and included every method that produced UNKNOWN calls, either natively or through a confidence threshold. The second was coverage calibration under shift, measured by coverage error and mean prediction-set size and computed only for methods that return prediction sets. The nominal miscoverage level was $\alpha = 0.10$, corresponding to a target coverage of $1 - \alpha = 0.90$.

Datasets.

- **Newborn human lung**: human, scRNA-seq; CZ CELLxGENE collection 6e00ccf7-0749-46ef-a999-dba785630d52; general and targeted (open set).
- **MacParland human liver**: human, scRNA-seq; GEO GSE115469; general and targeted (open set).
- **Healthy living-donor kidney**: human, scRNA/snRNA-seq; CZ CELLxGENE collection 0bae7ebf-eb54-46a6-be9a-3461cecefa4c; general and targeted (open set).
- **LungMAP broad-age lung**: human, snRNA-seq; CZ CELLxGENE collection 3de0ad6d-4378-4f62-b37b-ec0b75a50d94; general and targeted (fetal and child to adult shift).
- **Human retina bipolar**: human, scRNA/snRNA-seq; GEO GSE148077; general and targeted (fovea-to-periphery shift).

Comparators (method: family).

- scANVI[5]: variational autoencoder; UNKNOWN assigned when the maximum predicted class probability fell below the fifth percentile of that quantity among known calibration cells.
- scmap-cell[6]: nearest-neighbour/correlation mapping.
- scDOT[7]: optimal transport.
- mtANN[8]: multi-reference ensemble.
- Seurat label transfer[9]: anchor/CCA.
- CSA[10]: conformal annotation and the only other set-valued method in the comparison. On kidney, CSA accepted no known cells, so its accepted-known macro-F1 was undefined and was excluded from aggregation.

Observed evidence.

- **Targeted capability, novel-cell detection.** scOSCAR ranked fifth of ten on the novel-cell detection board (0.73), below Seurat label transfer, scANVI, mtANN, and scmap-cell. Against scANVI, UNKNOWN F1 was 0.844 versus 0.618 on newborn lung, 0.269 versus 0.311 on liver, and 0.279 versus 0.751 on kidney.
- **General performance, closed-set floor.** It ranked fourth on the closed-set board (0.82); known-cell macro-F1 was 0.936 versus 0.980 for the strongest comparator.
- **Targeted capability, calibration under shift.** Coverage error was approximately 0.03 and 0.01 at a mean prediction-set size of approximately 1.5. Point-classifier baselines returned no prediction sets. CSA over-rejected known cells under shift and scored 0.26 on the coverage composite, compared with 0.74 for scOSCAR.
- **Operating boundary.** Near-lineage novelty remained difficult: when a held-out type shared a retained sibling's manifold, UNKNOWN F1 fell to 0.279 versus 0.751 for scANVI on kidney.

CellWDA Selective, support-aware batch integration

Topic: Rare-cell-preserving batch integration

Problem (from the RPKG).
A batch-associated neighborhood may be a technical artifact, a counterpart missing from another batch or a real condition-specific state. Global batch removal cannot distinguish these cases and can over-correct rare or cohort-private cells.

Proposed Mechanism.
A count-aware negative-binomial variational-autoencoder latent, self-supervised conformal shared-support gating and support-weighted adversarial alignment, with a per-cell sharedness score and calibrated abstention set. *Grounded in: cell-weighted partial domain alignment.*

General performance.
Standard integration quality on cohorts in which most cell types were shared, scored on the scIB board. Metrics were isolated-label F1, adjusted Rand index, normalized mutual information, graph connectivity and isolated-label silhouette.

Targeted capability.
Rare-cell-aware integration: avoid over-mixing cells without a cross-batch counterpart while integrating shared populations. Metrics were shared-support AUROC, unshared false-mixing rate and integration local inverse Simpson's index; calibrated-abstention recovery of rare populations was an additional diagnostic.

Datasets.

- **Pancreas multi-study compendium**: human, scRNA-seq; scIB compendium (GSE84133, GSE85241, E-MTAB-5061, GSE81608, GSE86469 and GSE81076); general and targeted.
- **LungMAP broad-age lung**: human, snRNA-seq; CZ CELLxGENE collection 3de0ad6d-4378-4f62-b37b-ec0b75a50d94; general.
- **Kang PBMC (IFN-β)**: human, scRNA-seq; GEO GSE96583; general.
- **Liver (MacParland)**: human, scRNA-seq; GEO GSE115469; general and targeted.
- **Mouse hindbrain (developmental)**: mouse, scRNA-seq; Figshare 10.6084/m9.figshare.24625302; targeted.
- **Mouse retina bipolar (Shekhar)**: mouse, Drop-seq; GEO GSE81904; targeted.

Comparators (method: family).

- scVI[11]: variational autoencoder.
- Scanorama[12]: mutual-nearest-neighbour panorama stitching.
- Harmony[13]: iterative clustering and linear correction.
- scDML[14]: deep metric learning.
- BBKNN[15]: batch-balanced k-nearest neighbours.
- Seurat[9]: anchor/CCA.

Observed evidence.

- **Targeted capability, rare/private-state preservation.** CellWDA ranked first on the targeted board (0.78), with the highest mean shared-support AUROC among full-coverage methods (0.836 versus 0.809 for scVI and 0.754 for Scanorama). Its abstention set recovered 88.8% of documented rare cells in pancreas and retina versus approximately 80% for the scVI analogue.
- **General performance, integration quality.** It ranked first on the aggregate scIB board (0.64). Clustering performance was intermediate (mean isolated-label F1 of 0.409), whereas graph connectivity ranked first (0.968).
- **Operating boundary.** In the donor-only single-chemistry liver cohort, Scanorama recovered 98.9% of donor-confined cells versus 80.0% for CellWDA. In developmental-stage batches, low estimated cross-batch support made the calibrated abstention set large.

CONFIDE-VI Batch integration boundary case

Topic: Rare-cell-preserving batch integration

Problem (from the RPKG).
Local matched-pair or neighborhood-based correction can over-correct when cross-batch support is scarce or compositions differ, merging rare or batch-restricted states into abundant shared populations.

Proposed Mechanism.
A count-based negative-binomial warm-up latent, an FDR-controlled conformal alignability test of local cross-batch support, alignability-gated adversarial fine-tuning and a density-stratified Mondrian fidelity projection. *Grounded in: adversarial warm-up scheduling.*

General performance.
Standard batch-integration quality on discrete-batch cohorts, scored on the scIB board. Metrics were adjusted Rand index, integration and cell-type local inverse Simpson's indices, the k-nearest-neighbour batch-effect test and isolated-label F1.

Targeted capability.
Rare-cell alignability: whether the conformal alignability test separated true rare cells from batch artifacts. The primary metric was alignability AUROC; the rare-cell-scaffold composite and shared-support measures were secondary diagnostics.

Datasets.

- **Pancreas compendium**: human, scRNA-seq; GEO GSE84133 and E-MTAB-5061; general and targeted.
- **LungMAP broad-age lung**: human, snRNA-seq; CZ CELLxGENE collection 3de0ad6d-4378-4f62-b37b-ec0b75a50d94; general.
- **Mouse hindbrain**: mouse, scRNA-seq; Figshare 10.6084/m9.figshare.24625302; targeted.
- **Kang PBMC (IFN-β)**: human, scRNA-seq; GEO GSE96583; general and targeted.
- **CellBench cell-line mixture**: human cell lines, scRNA-seq; GEO GSE118767; general.
- **PBMC-SCA**: human, scRNA-seq; GEO GSE132044; targeted.

Comparators (method: family).

- scVI[11]: variational autoencoder.
- Scanorama[12]: mutual-nearest-neighbour panorama stitching.
- Harmony[13]: iterative clustering and linear correction.
- scDML[14]: deep metric learning.
- STACAS[16]: Seurat-anchor-based integration.
- Seurat[9]: anchor/CCA.

Observed evidence.

- **General performance, integration quality.** CONFIDE-VI ranked fourth on the scIB board (0.57). Its adjusted Rand index was comparable with the strongest active comparators, while batch mixing was deliberately conservative (the lowest integration local inverse Simpson's index).
- **Secondary scaffold behavior.** Low-density stratum-0 cells retained more rare-marker signal than higher-density cells of the same type (ratio 1.665 in pancreas and 1.057 in PBMC-SCA), whereas the comparator analogues were below one.
- **Primary targeted claim not supported.** The conformal alignability score separated true rare cells only near chance (mean AUROC 0.47). Aggregate corrected-layer marker fidelity was also not supported: top-k rare-marker recovery was lower by 0.073 and log-fold-change sign concordance was lower by 0.309.
- **Operating boundary.** Cohorts whose batch axis encoded continuous development fell outside the intended scope.

Supplementary Note 3: Secondary blinded model ratings of candidate formulations

The blinded human comparison in the main text measured which of three candidates a researcher would choose to investigate further. We also scored the written candidates with two language-model judges under a fixed rubric. For each of the same 30 Problems, the scHypothesizer, Abstract-only and Problem-only candidates were anonymized and order-shuffled. Two language-model judges independently scored every candidate three times on novelty, significance, mechanistic specificity, biological grounding and feasibility, using an integer scale from 1 to 10. The resulting six ratings per candidate were summarized by topic and dimension (Supplementary Table S4).

Across all topics and dimensions, the mean scores were 7.49 for scHypothesizer, 6.72 for Abstract-only and 6.48 for Problem-only. The largest separations occurred for novelty and mechanistic specificity, whereas significance, biological grounding and feasibility were more similar across conditions. These scores depend on the rubric and judge models and were not used to select candidates.

Table S4 | Secondary blinded model ratings by topic and dimension. Values are means over the ten candidate triplets in each topic and six blinded ratings per candidate (two judge models, each applied three times; integer scale, 1–10). Bold indicates the highest mean within a row. The rubric is given in Supplementary Table S5.

Dimension	scHypothesizer	Abstract-only	Problem-only
Ambient decontamination			
Novelty	**7.28**	5.12	4.13
Significance	**7.32**	6.92	6.50
Mechanistic specificity	**8.18**	6.92	6.57
Biological grounding	**7.73**	7.58	7.40
Feasibility	**7.12**	6.50	6.67
Rare-cell integration			
Novelty	**7.62**	5.83	5.30
Significance	7.03	**7.17**	6.87
Mechanistic specificity	**8.30**	7.10	6.97
Biological grounding	6.50	**6.68**	6.67
Feasibility	**6.95**	6.23	6.70
Reference mapping with calibrated uncertainty			
Novelty	**7.97**	6.35	5.47
Significance	**7.33**	7.25	6.93
Mechanistic specificity	**8.50**	7.55	7.17
Biological grounding	**7.47**	6.93	6.77
Feasibility	**7.10**	6.67	7.08
All topics pooled			
Novelty	**7.62**	5.77	4.97
Significance	**7.23**	7.11	6.77
Mechanistic specificity	**8.33**	7.19	6.90
Biological grounding	**7.23**	7.07	6.94
Feasibility	**7.06**	6.47	6.82

Table S5 | Rubric for the secondary blinded model ratings. Each candidate was scored on five dimensions on an integer scale from 1 to 10, with anchor descriptions at five score bands.

Score	Anchor description
Novelty: distinctness from nearby existing approaches for the Problem	
9–10	Strikingly original Mechanism, with no clear precedent for the Problem.
7–8	Clearly distinct, with a genuinely new combination or technical change.
5–6	A recognizable approach containing a non-trivial new element.
3–4	A minor variation on a well-known approach.
1–2	A standard or textbook approach without a new Mechanism.
Significance: expected advance for the Problem if the proposal works	
9–10	Would substantially advance or redefine the Problem.
7–8	Would provide a clear and important advance.
5–6	Would provide a moderate, incremental advance.
3–4	Would have marginal importance even if successful.
1–2	Would have negligible impact or target a non-problem.
Mechanistic specificity: concreteness of components, design choices and actionable steps	
9–10	Fully concrete and complete; implementable as written.
7–8	Mostly concrete, with only minor under-specification.
5–6	Partly specified, with several important design choices left open.
3–4	Mostly generic, with few concrete choices.
1–2	A vague description without an actionable Mechanism.
Biological grounding: grounding in scRNA-seq biology and data properties	
9–10	Deeply and specifically grounded in the relevant biology and data properties.
7–8	Well grounded and clearly uses relevant biological or data structure.
5–6	Partly grounded, but includes a substantially generic machine-learning component.
3–4	Largely generic and mostly ignores relevant data or biological properties.
1–2	Ignores or contradicts the biology or data properties of the Problem.
Feasibility: implementability and benchmarkability with available tools and data	
9–10	Directly implementable with standard tools and ready benchmarks.
7–8	Feasible with moderate effort; data and metrics are clearly identifiable.
5–6	Feasible, but with notable engineering or data challenges.
3–4	Substantial feasibility concerns or difficult data and benchmarking requirements.
1–2	Impractical or not benchmarkable as described.

Matched comparison of simple RAG and pre-critic MCG

We also applied the same rubric to the matched simple RAG and pre-critic MCG candidates described in the main text. These two conditions used the same Problem, language-model backend, output schema and corpus of complete novelty summaries; they differed in whether evidence was selected by title similarity or through the typed graph pairing and organized around the retrieved Mechanism. Their pooled mean scores were similar (6.80 and 6.87, respectively), even though the masked structural audit distinguished their organization. The final scHypothesizer records are included for descriptive context but were the only condition to undergo the subsequent Proposer-Critic loop. Consequently, their higher pooled mean of 7.24 cannot be attributed to MCG's evidence-selection and organization step alone (Supplementary Fig. S1 and Supplementary Table S6).

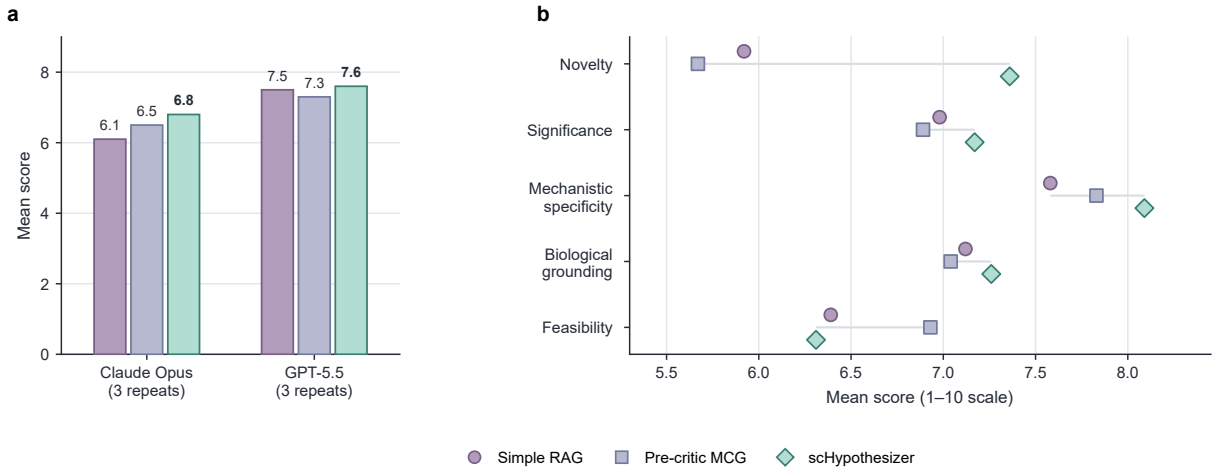


Figure S1 | Detailed secondary model ratings for the matched construction comparison. a, Mean score from each blinded judge model, with each judge applied three times per candidate. **b**, Pooled mean for each of the five rubric dimensions. Simple RAG and pre-critic MCG used a matched corpus and generation protocol but selected and organized evidence differently. The final scHypothesizer records additionally underwent Proposer-Critic refinement.

Table S6 | Model ratings in the matched comparison by topic and dimension. Simple RAG and pre-critic MCG used the same source corpus and generation backend but selected and organized evidence differently. Final scHypothesizer records are shown for context; because only that condition underwent Proposer-Critic refinement, comparisons with the final records do not isolate the contribution of MCG's evidence-selection and organization step.

Dimension	scHypothesizer	Pre-critic MCG	Simple RAG
Ambient decontamination			
Novelty	**7.55**	6.05	5.90
Significance	**7.43**	7.30	7.08
Mechanistic specificity	**8.33**	8.10	7.78
Biological grounding	**8.42**	8.20	8.02
Feasibility	6.37	**6.90**	6.68
Mean of five dimensions	**7.62**	7.31	7.09
Rare-cell integration			
Novelty	**7.17**	5.00	6.08
Significance	6.88	6.47	**7.00**
Mechanistic specificity	**7.83**	7.55	7.57
Biological grounding	**6.55**	6.15	6.52
Feasibility	6.08	**7.03**	5.85
Mean of five dimensions	**6.90**	6.44	6.60
Reference mapping with calibrated uncertainty			
Novelty	**7.37**	5.97	5.77
Significance	**7.18**	6.90	6.87
Mechanistic specificity	**8.12**	7.83	7.40
Biological grounding	6.82	6.77	**6.83**
Feasibility	6.48	**6.87**	6.65
Mean of five dimensions	**7.19**	6.87	6.70
All topics pooled			
Novelty	**7.36**	5.67	5.92
Significance	**7.17**	6.89	6.98
Mechanistic specificity	**8.09**	7.83	7.58
Biological grounding	**7.26**	7.04	7.12
Feasibility	6.31	**6.93**	6.39
Mean of five dimensions	**7.24**	6.87	6.80

Supplementary Note 4: Catalog of consolidated method hypotheses

This catalog records the 30 consolidated candidate outputs considered in the hypothesis-generation analysis. The cards are titled Ambient Hypothesis 1–10, Integration Hypothesis 1–10, and Reference Mapping Hypothesis 1–10. Card text is reproduced from the system output and condensed for readability; the complete structured records are provided in Supplementary Data 1, where they appear in the same order.

Ambient decontamination

Ambient Hypothesis 1

Problem. Droplet single-cell RNA sequencing often contains ambient transcripts that create weak ectopic expression of genes from abundant populations in unrelated cells. General-purpose ambient correction can be insufficient for within-cell-type differential expression: conservative correction may leave ambient-derived genes significant, whereas aggressive correction may perturb many unrelated genes or erase rare-cell markers. The task is, for each sample, cell, and gene in a raw count matrix, to separate observed molecules into native expression versus sample-specific ambient contribution, producing decontaminated counts and per-cell contamination burdens that preserve rare-cell programs while reducing false marker discovery and false-positive within-cell-type DGE.
Proposed mechanism. The method models each sample separately and treats the same gene as potentially native in one cell neighborhood but ambient-derived elsewhere. It first estimates a sample-specific ambient profile from empty droplets or low-count barcodes when available, with fallback estimation from low-complexity cells flagged for audit. It then constructs a mutual nearest-neighbor cell graph using features depleted for strong ambient enrichment, so that rare populations can obtain local support without requiring hard clusters. For each gene, observed counts are fit with an additive count model whose intensity is the sum of a native component and a cell exposure-scaled ambient component. The native component is regularized by graph total variation, encouraging compact locally coherent expression islands while allowing sharp boundaries. Final corrected counts are obtained by posterior allocation of observed molecules between native and ambient intensities, accompanied by gene-, cell-, and sample-level removal summaries for DGE interpretation.
Key steps.

- Estimate a per-sample ambient gene profile and cell-specific exposure score from empty or low-count droplets.
- Build an ambient-resistant mutual nearest-neighbor graph within each sample using ambient-depleted expression features.
- Fit an additive count decomposition into graph-regularized native expression and exposure-scaled ambient signal for every cell-gene pair.
- Allocate observed molecules probabilistically to native versus ambient components to generate corrected count-like matrices.
- Report correction audits including removed counts by cell, gene, sample, and their relationship to downstream DGE signals.

Ambient Hypothesis 2

Problem. Droplet-based single-cell RNA sequencing can convert sample-specific extracellular RNA into false-positive subpopulation differentially expressed genes. Ambient contaminant genes may be ranked as significant within cell subpopulations, such as immunoglobulin genes in T-cell subsets or hemoglobin genes in fetal liver B-cell subsets, even when those cells do not endogenously express them. Given raw UMI counts with sample, batch, experimental-group, and cell-type or preliminary cluster labels, plus empty-droplet or low-UMI ambient libraries per sample, the task is to generate a corrected count layer that removes sample-specific ambient RNA from non-expressing subpopulations while preserving marker expression in rare populations near 1-2% abundance.
Proposed mechanism. The method performs context-aware ambient correction rather than applying one global subtraction rule per gene or per cell. First, it estimates a sample-specific ambient expression direction from empty-droplet or low-UMI libraries and restricts correction to genes recurrently observed in those ambient profiles. Second, using preliminary clusters or annotations, it identifies rare-population marker genes and creates a protected gene-by-cluster mask: an ambient-prone gene can be shielded in the cluster where it is an endogenous marker but remain correctable elsewhere. Third, for each sample-by-cluster context, the method builds an eligible ambient vector after excluding protected markers active in that context. Cell count vectors are then projected away from this context-specific ambient direction and reconstructed as nonnegative corrected counts anchored to the cluster's endogenous expression profile. Genes not supported by the ambient recurrence set are passed through unchanged. This yields a corrected layer compatible with standard clustering, annotation, and pseudobulk differential-expression workflows.
Key steps.

- Estimate per-sample ambient profiles from empty-droplet or low-UMI libraries and define recurrent ambient-prone genes.
- Identify rare cluster-enriched markers and create protected gene-by-cluster shields for endogenous rare-population signal.
- For each sample and cluster, construct an eligible ambient direction excluding protected markers for that context.
- Project cell counts away from the eligible ambient direction under nonnegativity and cluster-profile anchoring constraints.
- Return corrected counts and diagnostics separating contaminant-gene reduction from protected-marker retention.

Ambient Hypothesis 3

Problem. Ambient RNA contamination in droplet single-cell RNA sequencing is experiment-specific: both the contaminating transcript composition and its magnitude vary across samples. Therefore, fixed or generic background correction can fail in complex tissues, necrotic material, tumor biopsies, and erythrocyte-lysed preparations. The task is to accept a count matrix with raw UMI counts, sample identifiers, optional empty-droplet or low-UMI barcodes, and optional labels for evaluation, then infer sample-level ambient profiles, per-cell contamination fractions, gene-by-cell ambient probabilities, corrected counts, and analysis-ready embeddings or clusters. The key failure mode is removing leaked marker transcripts from marker-negative cells while not erasing the same genes when they represent native expression in rare cell populations.
Proposed mechanism. The method estimates an ambient transcript distribution separately for each sample, using empty or low-UMI droplets when available and a regularized posterior to avoid unstable zero estimates. A non-neural multinomial mixture then decomposes each cell's observed counts into native and ambient components, yielding cell-specific contamination fractions and initial gene-by-cell ambient probabilities. To handle rare-marker collisions, each expressed cell-gene pair receives a local native-support statistic: observed expression in a neighborhood of transcriptionally similar cells is compared with the expression expected from the sample ambient profile and the cell's contamination fraction. A monotone family of correction rules protects genes with increasing local support from subtraction while removing unsupported ambient-like counts. Split calibration examples, built from high-support marker-like pairs with controlled synthetic ambient additions, are used to select the most aggressive correction rule whose upper bound on rare-marker under-retention risk remains below a user-chosen tolerance.
Key steps.

- Estimate a regularized ambient profile independently within each sample from empty or low-count droplets.
- Fit a multinomial native-plus-ambient mixture to infer per-cell contamination fractions and baseline ambient probabilities.
- Compute neighborhood native-support scores for each observed cell-gene pair relative to expected ambient contribution.
- Calibrate the protection strength using split risk control on rare-marker collision examples with added ambient signal.
- Produce corrected counts, ambient probability matrices, contamination summaries, embeddings, and clusters.

Ambient Hypothesis 4

Problem. Droplet-based single-cell RNA-seq preprocessing can fail when contamination fractions are estimated inaccurately or when users intentionally prefer stronger background removal than a point estimate would imply. The task is to start from raw integer UMI counts, batch labels, and optional cluster or cell-type annotations, and infer both native and contaminant count matrices. The central failure mode is that ambiguous low-count marker UMIs may be background in unrelated cells but true expression in rare populations. The method must support deliberate overcorrection to remove background signal while avoiding elimination of plausible native marker expression, and should quantify uncertainty in cell-wise contamination.
Proposed mechanism. Model ambient removal as posterior count allocation followed by risk-aware integer subtraction. For each batch, estimate a background expression profile from empty or low-UMI droplets when available, otherwise from low-confidence called-cell signal. Assign cells to populations using supplied annotations or unsupervised clustering. Each observed cell-gene count is represented as a mixture of native population expression and contamination, where contamination combines batch-specific background with expression contributed by other populations. A variational expectation-maximization procedure estimates native profiles, contaminant profiles, cell-specific contamination fractions, posterior contaminant-count distributions, and credible intervals. Rare populations are handled with a prevalence-adaptive empirical prior that stabilizes only gene-population pairs showing enrichment evidence, rather than protecting all genes in small groups. Final subtraction uses an asymmetric posterior-quantile decision rule: ambiguous counts are removed more aggressively when background-removal risk dominates, but over-subtraction penalties are increased for enriched markers in rare populations.
Key steps.

- Estimate batch-specific ambient profiles from empty, low-UMI, or low-confidence background droplets.
- Define cell populations from provided labels or clustering and fit a native-contamination count mixture.
- Infer posterior contaminant counts and cell-wise contamination fractions with credible intervals.
- Apply prevalence-adaptive shrinkage to stabilize rare enriched population-marker expression.
- Subtract integer posterior quantiles under asymmetric loss and output native, removed, and diagnostic matrices.

Ambient Hypothesis 5

Problem. Droplet single-cell RNA sequencing analyses for within-cell-type differential gene expression can fail when ambient RNA correction is too global or insufficiently selective. General-purpose decontamination may leave ambient-derived genes falsely significant within annotated cell types, yet aggressive matrix-wide subtraction can perturb many unrelated genes and remove weak endogenous markers from rare populations. The task is to start from a raw or near-raw count matrix, with cell-containing droplets and empty or low-UMI droplets when available, estimate sample-specific ambient profiles and per-cell contamination burden, identify globally contaminating genes, and produce corrected counts in which only statistically supported contaminating gene-cell entries are edited while non-contaminating genes and protected rare-marker entries remain exactly unchanged.
Proposed mechanism. The method would fit a sample-specific endogenous-plus-ambient count model that separates cluster-level biological expression from cell-level contamination exposure. Empty droplets, or low-UMI droplets when empties are unavailable, define an ambient multinomial profile for each sample or batch. Cell labels may be supplied or estimated by conservative clustering, after which a negative-binomial observation model estimates cluster-specific endogenous expression, gene dispersion, and per-cell ambient load. Genes are called globally contaminating only when they are enriched in the ambient profile, appear broadly outside plausible source populations, and show residual dependence on cell contamination load. For each gene-cell entry, an ambientness score is computed from the fitted model. A conformal rare-marker shield calibrates edit thresholds using high-confidence endogenous marker counts from rare or small populations, so marker-like entries are protected unless their evidence for contamination is extreme. For editable entries only, posterior expected ambient counts are removed by conservative thinning; all other counts are copied exactly.
Key steps.

- Estimate per-sample ambient RNA composition from empty or low-UMI droplets and infer per-cell contamination load.
- Fit a cluster-aware endogenous-plus-ambient count model with gene dispersion and cluster-specific expression.
- Call globally contaminating genes using ambient enrichment, broad ectopic distribution, and association with contamination load.
- Calibrate conformal edit thresholds on high-confidence rare-population marker entries to protect endogenous low-count signals.
- Apply selective count thinning only to eligible contaminating gene-cell entries and emit corrected counts, edit masks, contamination scores, and gene calls.

Ambient Hypothesis 6

Problem. Human heart snRNA-seq contains severe cytoplasmic and cell-free RNA contamination, often dominated by cardiac muscle transcripts, that masks cell-type-specific nuclear signal, blurs the distinction between empty droplets and nuclei-containing droplets, biases nuclei calling, and distorts downstream interpretation. The task is to use the raw droplet-level count matrix, including empty libraries, contaminated droplets, and putative nuclei, to select a sample-specific empty-library UMI threshold for ambient RNA estimation. The method must then generate corrected counts that reduce ectopic cardiac muscle ambient signal while preserving marker expression from rare cardiac nuclear populations.
Proposed mechanism. The method treats empty-library cutoff selection as a maximum tolerated threshold problem. For each sample, it sweeps candidate low-UMI thresholds and estimates an ambient profile from droplets below each cutoff. At each candidate cutoff, it quantifies whether coherent low-prevalence cardiac nuclear marker programs appear enriched in the proposed empty-library pool, indicating that true rare nuclei are beginning to contaminate the ambient estimate. A monotone risk curve is fitted across increasing thresholds, and the selected cutoff is the largest one whose estimated rare-marker leakage probability remains below a prespecified safety bound. The resulting ambient profile is used for per-nucleus expected ambient count estimation. Subtraction is expression-protected: genes are removed when consistent with ambient contamination, but subtraction is capped when the gene is locally enriched among neighboring nuclei with similar expression, preserving true cardiac muscle or rare-population signal. The method outputs the chosen threshold, ambient profile, contamination scores, and a decontaminated count layer.
Key steps.

- Identify conservative high-confidence nuclei and derive coherent low-prevalence cardiac marker programs.
- Sweep candidate empty-library UMI thresholds and estimate an ambient RNA profile for each threshold.
- Measure rare-program leakage into each putative empty-library pool and fit a monotone leakage-risk curve.
- Select the highest threshold with leakage risk below the user-defined safety bound.
- Subtract expected ambient counts per nucleus while protecting locally enriched marker expression.

Ambient Hypothesis 7

Problem. Ambient RNA decontamination in droplet single-cell RNA sequencing can fail by overcorrecting harmless ambient genes. Correcting every detected ambient-compatible transcript can unnecessarily perturb count matrices, change differentially expressed gene identities, and remove true low-level expression, especially in rare populations. The task is to produce a selectively decontaminated count matrix from counts, sample labels, cell or cluster annotations, and optionally empty or low-UMI droplets, removing ambient RNA likely to alter marker, rare-cluster, or pseudo-bulk differential expression decisions while preserving rare-cell marker-supported expression and ambient-compatible counts that are too infrequent to affect testing.
Proposed mechanism. The method estimates a sample-specific ambient expression profile and cell-level exposure, but uses these estimates only to define cell-gene upper bounds on plausible removable counts rather than subtracting an expected value. It then computes local endogenous-support scores from neighborhoods or annotated clusters, so genes coherently expressed in rare cells receive lower editability and higher removal cost. A block-wise sparse integer optimization is solved to find the smallest count edits within ambient-plausible bounds that satisfy decision-margin constraints. These constraints preserve supported rare-cell marker margins, reduce ectopic marker signals in nonexpressing groups, and repair pseudo-bulk differential expression changes attributable to ambient burden. Counts that are ambient-compatible but do not threaten marker ranking, cluster identity, or differential expression margins remain unchanged by optimality. The output includes corrected counts plus per-gene, per-cell, or per-cell-type decision-impact scores suitable for standard clustering, marker testing, and pseudo-bulk analysis.
Key steps.

- Estimate sample-specific ambient profiles and cell-level exposure from empty, low-UMI, or background-enriched droplets.
- Convert expected ambient burden into integer upper bounds for removable counts at each cell-gene pair.
- Compute local endogenous-support scores to protect coherent low-level expression in rare neighborhoods or clusters.
- Solve a sparse minimal-edit projection that enforces marker-preservation and ambient-risk repair margins.
- Return corrected counts and decision-impact scores for genes, cells, and annotated groups.

Ambient Hypothesis 8

Problem. Droplet-based single-cell RNA sequencing datasets can exhibit mixed-severity ambient contamination within each biological sample. Some transcripts are highly contaminating and are best handled by targeted correction, while many others create weak, diffuse contamination that targeted contaminating-gene detection may miss. This creates a difficult setting in which the same ambient-enriched gene may represent leakage across unrelated cells or true endogenous expression in a rare, coherent cell population. The task is to estimate sample-specific ambient profiles and produce corrected expression values that remove both severe and low-level ambient RNA while preserving rare-cell marker signal and reporting interpretable evidence for limited correction.
Proposed mechanism. The method estimates an ambient profile separately for each sample using empty droplets when available, or low-UMI libraries otherwise, and converts each sample-gene pair into a continuous ambient-risk score rather than a binary contaminant label. It builds a cell-neighborhood graph from features down-weighted for ambient risk, reducing the chance that leakage itself defines neighborhoods. For each sample and ambient-risk gene, it specifies an ambient-only null model using the sample ambient profile, cell library size, and an estimated cell-specific contamination load. Local connected graph regions are then tested for expression exceeding this null, with false-discovery control applied at the connected-support level. Significant supports are annotated as protected endogenous regions. Correction is performed by gated ambient thinning: counts are removed strongly outside protected supports and conservatively inside them, allowing a gene to be corrected in unrelated cells but retained in rare coherent populations. The output includes corrected counts, ambient profiles, contamination estimates, and protected-support annotations.
Key steps.

- Estimate per-sample ambient RNA profiles and cell-specific contamination loads from empty droplets or low-UMI libraries.
- Assign continuous ambient-risk scores to sample-gene pairs and construct an ambient-masked cell-neighborhood graph.
- For each ambient-risk gene, test connected graph neighborhoods against a sample-specific ambient-only count null.
- Declare false-discovery-controlled protected supports where expression is statistically coherent beyond ambient expectation.
- Apply support-aware ambient thinning to produce corrected expression while recording correction strength and protection status.

Ambient Hypothesis 9

Problem. Droplet single-cell RNA sequencing preprocessing requires estimating ambient RNA contamination fractions from raw UMI counts, yet these estimates can be inaccurate across datasets and sensitive to normalization, clustering, or rare-cell prevalence. The task is to infer per-cell contamination fractions and produce a corrected raw-count matrix from cell-containing barcodes, optionally using empty or low-count barcodes, while allowing deliberate overcorrection when needed to remove background signal. The failure mode is that ambient transcripts from abundant populations can create false marker expression in other cells, whereas aggressive correction can erase true marker UMIs from rare populations whose native expression is sparsely observed.
Proposed mechanism. The method operates directly on raw UMI counts and treats each observed gene-cell count as a mixture of native molecules and ambient molecules. Empty or low-count droplets are split into fitting and calibration subsets to estimate the sample-specific ambient RNA profile and derive calibrated gene-wise upper envelopes for counts plausibly generated by ambient RNA at different burden levels. Per-cell contamination fractions are then estimated by a raw-count likelihood that uses library size and ambient profile but discounts counts exceeding the calibrated ambient envelope, so rare-cell marker counts do not artificially increase the inferred contamination rate. Finally, each gene-cell count is probabilistically allocated between ambient and native components, with removable ambient counts capped by the calibrated envelope. Counts above the envelope are protected as native evidence, yielding corrected raw counts, per-cell contamination estimates, and allocation uncertainty in standard annotated single-cell matrix outputs.
Key steps.

- Estimate a sample-specific ambient expression profile from empty or low-count droplets using raw UMI counts.
- Construct split-calibrated per-gene ambient count envelopes across candidate ambient burdens.
- Estimate each cell's contamination fraction with a raw-count likelihood constrained by ambient-plausibility envelopes.
- Allocate observed UMIs into ambient-compatible and protected-native components using capped posterior expectations.
- Return corrected counts, contamination scores, and allocation uncertainty for downstream analysis.

Ambient Hypothesis 10

Problem. Ambient RNA contamination in droplet single-cell RNA sequencing can create false-positive subpopulation differentially expressed genes: contaminant transcripts are ranked as significant markers within rare or fine-grained subclusters despite lacking endogenous support. Examples include immunoglobulin transcripts in T-cell subpopulations or hemoglobin transcripts in non-erythroid B-cell subsets. The task setting includes count matrices with sample labels, parent cell-type or cluster labels, rare-subcluster labels, and optionally empty or low-UMI droplets. The goal is to produce a corrected count layer and an interpretable gene-by-subpopulation decision table that removes ambient-explained marker calls while preserving true rare markers, neighborhoods, clustering, and annotation.
Proposed mechanism. The method estimates a sample-specific ambient transcript profile from empty or low-UMI droplets and converts it into a conservative upper bound on removable counts for every cell-gene pair. Within each parent cell type or cluster, it performs rare-subpopulation marker testing, then reframes each candidate marker as a counterfactual decision: would subtracting only ambient-plausible counts be sufficient to remove marker status by rank, log fold change, or adjusted significance? For each gene-subpopulation pair, it searches for the smallest ambient-feasible edit concentrated in the relevant cells. If that edit flips the marker decision, the counts are minimally corrected and the pair is labeled ambient-explained. If the marker remains significant even after maximal plausible ambient removal, it is protected as endogenous-supported. Boundary cases with unstable decisions are labeled ambiguous rather than forced. All components are directly tied to ambient-profile estimation, counterfactual marker flipping, and context-specific marker protection.
Key steps.

- Estimate per-sample ambient gene probabilities from empty or low-UMI droplets.
- Define a conservative ambient-feasible removable count budget for each observed cell-gene entry.
- Test markers for each rare subpopulation within its parent cell type or cluster.
- For each candidate marker, compute the minimal ambient-feasible count edit needed to lose marker status.
- Output corrected counts and classify each gene-subpopulation pair as ambient-explained, protected, or ambiguous.

Batch integration

Integration Hypothesis 1

Problem. In single-cell RNA-seq batch integration, overmixing reduces biological fidelity: embeddings may mix technical batches well yet blur cell-type boundaries and erase subtle, biologically meaningful structure. The task is to learn an integrated latent representation from annotated count matrices, with raw counts available in the main matrix or count layer, required batch annotations, and optional cell-type annotations used only for evaluation. The embedding should mix batches for shared populations while preserving local structure and separability of rare or cohort-private states. The key unsupervised difficulty is distinguishing cells with credible cross-batch counterparts from cells that are compositionally unmatched and should be protected from overcorrection.
Proposed mechanism. The method uses calibrated shared-support gating to modulate adversarial batch alignment. It first trains a compact count-aware variational autoencoder with a negative-binomial likelihood on highly variable genes, yielding a denoised warm-start latent space while retaining count structure. In that space, it forms self-supervised rare-state exclusion episodes by clustering cells and temporarily treating selected pseudo-populations as absent from selected target batches, creating calibration examples of cells that should not be considered shared. For each cell and candidate target batch, a density-normalized nearest-neighbor nonconformity score compares distance to target-batch neighbors with the local same-batch distance scale. Split-conformal calibration sets thresholds that bound false shared-support calls at a chosen risk level. The inferred per-cell support set is transformed into a detached weight on a gradient-reversal batch-discriminator loss. Final training jointly optimizes count reconstruction, latent regularization, and support-gated batch removal, applying strong alignment only where calibrated cross-batch support exists.
Key steps.

- Train a negative-binomial latent count model on highly variable genes to obtain a denoised initial embedding.
- Create self-supervised calibration episodes by withholding clustered pseudo-populations from target batches to mimic private states.
- Compute density-normalized cross-batch neighborhood nonconformity scores for each cell against each candidate target batch.
- Use split-conformal thresholds to infer calibrated shared-support sets and convert them into detached adversarial weights.
- Retrain or fine-tune the latent model with reconstruction, regularization, and support-gated batch-adversarial objectives.

Integration Hypothesis 2

Problem. In unsupervised integration of single-cell RNA count datasets with batch annotations but cell-type labels held out for evaluation, existing methods often assume substantial shared cell-type composition across batches. Under limited cell overlap, especially when cohort sizes and compositions are imbalanced, they may neither align truly shared populations nor protect cells present in only one or a few batches. The concrete failure mode is an embedding that appears well mixed globally yet incorrectly merges biologically distinct or batch-specific cell types while shared types remain distorted or inconsistently aligned. The task is to learn a latent representation from counts and batch labels that separates technical effects from genuine partial-overlap biology.
Proposed mechanism. The method trains a compact count-aware variational autoencoder on highly variable genes. Its encoder does not receive batch identity, but every encoder normalization layer computes hidden activation means and second moments separately within each batch represented in the minibatch, then normalizes using an equal-weight average of those per-batch moments. Running statistics are updated with the same equal-batch rule, so the coordinate system is calibrated to a batch-balanced distribution rather than a cell-count-weighted distribution dominated by large cohorts or abundant shared types. The decoder receives the latent variable plus a learned batch embedding and reconstructs raw counts with a negative-binomial likelihood, allowing technical count shifts to be modeled outside the latent code. Training uses only counts and batch labels. After fitting, weights are frozen and normalization statistics are recalibrated under proportional versus equal-batch loaders; changes in rare-cell neighborhoods quantify whether normalization statistics are driving overcorrection.
Key steps.

- Select highly variable genes from raw counts and train using counts plus batch labels only.
- Encode cells with batch-free layers whose normalization moments are averaged equally across batches.
- Decode latent variables with batch embeddings under a negative-binomial count likelihood.
- Export the fitted latent means as the integrated embedding.
- Recalibrate only encoder normalization statistics under alternative loaders to test rare-neighborhood sensitivity.

Integration Hypothesis 3

Problem. Large single-cell RNA-seq studies often combine donors, tissues, protocols, or disease cohorts whose cell-state composition is not shared across batches. Local matched-pair or neighborhood-based correction can become unstable when cross-batch matches are scarce or compositions differ, improving global mixing while erroneously merging batch-specific cell types or rare biological states into abundant shared populations. The task is to learn an integrated embedding from raw counts and batch labels, with optional cell-type labels reserved only for evaluation, that removes technical variation among genuinely shared states while preserving rare or batch-restricted neighborhoods, without using cell-type labels or cross-batch anchors during training.
Proposed mechanism. The method first learns a reconstruction-driven latent manifold from counts using a negative-binomial variational model with a batch-conditioned decoder and library-size covariate, before applying any batch-removal pressure. This warm-up stage is intended to stabilize local transcriptomic topology, especially for rare neighborhoods. The latent space is then partitioned into small local neighborhoods. For each neighborhood and source batch, the algorithm computes density-normalized affinity to candidate neighborhoods in other batches and evaluates whether the apparent counterpart is stronger than null neighborhoods matched for local density and batch size. Conformal or target-decoy calibration yields p-values, followed by false-discovery-rate control across all local alignability decisions. During fine-tuning, a batch discriminator with gradient reversal is applied only to cells in neighborhoods declared alignable. Neighborhoods without statistically supported cross-batch counterparts remain governed mainly by count reconstruction and are marked as protected batch-restricted or rare states.
Key steps.

- Train a count-based latent model with batch-aware decoding and library-size adjustment without batch-invariance pressure.
- Partition the warm-up latent manifold into compact neighborhoods and estimate density-normalized cross-batch affinities.
- Calibrate local alignability tests using matched null neighborhoods and control the false discovery rate over all tests.
- Fine-tune the latent model with batch-removal gradients only in neighborhoods passing the alignability criterion.
- Export the final embedding and per-neighborhood protection status for downstream evaluation and interpretation.

Integration Hypothesis 4

Problem. Integrated single-cell RNA-seq analyses often provide a batch-mixed representation but lack corrected expression matrices that can be used for marker-gene inspection and differential-expression-style analysis after integration. The task is to use raw counts, batch labels, and optional cell annotations reserved for evaluation to learn both an integrated latent representation and a batch-corrected gene-expression layer. The correction must remove technical variation across heterogeneous cohorts while preserving biologically meaningful within-batch gene-level contrasts, especially rare-state versus nearby parent-lineage log fold-changes whose marker ranking, sign, and magnitude can be distorted by aggressive correction.
Proposed mechanism. The method uses a lightweight count-based variational model with an encoder that maps log-normalized highly variable gene profiles to a batch-agnostic biological latent state. A decoder produces a canonical corrected expression composition in log-ratio or simplex coordinates, while separate batch-specific gene perturbations are used only to reconstruct observed counts through an overdispersed count likelihood. This separates the corrected matrix from nuisance batch shifts without forcing batch information into the biological embedding. A held-out calibration split estimates conformal tolerances for distortion of same-batch local log fold-change contrasts in compositional geometry. Calibration is stratified by local expression-density bands so rare-like cells receive tolerances estimated from similarly sparse neighborhoods rather than from abundant populations. At inference, decoded corrected expression is minimally projected or adjusted to satisfy the calibrated local contrast-fidelity constraints, yielding both an integrated embedding and a corrected expression layer.
Key steps.

- Encode normalized gene profiles into a batch-agnostic latent representation while retaining raw counts for likelihood-based training.
- Decode latent states into canonical corrected compositional expression and model observed counts using batch-specific gene perturbations only in the reconstruction path.
- Build held-out same-batch local micro-contrasts between rare-like cells and nearby parent-like neighborhoods in compositional log-ratio space.
- Calibrate density-stratified conformal thresholds for local log fold-change distortion.
- Project decoded corrected matrices to the nearest values satisfying calibrated contrast-fidelity constraints and export the embedding and corrected layer.

Integration Hypothesis 5

Problem. Heterogeneous single-cell RNA-seq cohort integration often yields useful low-dimensional embeddings but not batch-corrected gene-expression matrices suitable for marker-gene identification and differential-expression analysis after integration. The task is to take count matrices with batch labels, shared highly variable genes, and optional annotations for evaluation, and learn both an integrated representation and corrected gene-level values. The difficult setting includes unequal cell-type composition, rare or cohort-enriched states, and protocol-specific gene distortions. Shared abundant populations should mix after correction, while rare or batch-restricted populations should remain transcriptionally distinct and retain marker programs.
Proposed mechanism. Model normalized log-expression as a shared nonnegative biological expression profile observed through a low-capacity, batch-specific gene-wise measurement transformation. Each cell is represented by nonnegative latent factors, decoded through a shared nonnegative gene-signature matrix to produce the corrected expression vector. The observed expression in each batch is reconstructed by applying a batch-specific gene offset and positive gene scale to that shared biological profile. Batch parameters are estimated conservatively using a worst-stratum reconstruction objective over latent factor-defined subgroups within each batch, rather than average reconstruction error across all cells. This makes rare or high-error subpopulations influence calibration instead of being overwhelmed by abundant shared cell types. The loss combines robust reconstruction, worst-stratum weighting, sparsity on factors and signatures, and penalties on offset and scale magnitude to prevent the batch model from absorbing genuine biology. Corrected expression is the shared decoded profile, and the integrated representation is the nonnegative factor vector.
Key steps.

- Normalize counts, select shared variable genes, and initialize nonnegative cell factors and gene signatures from pooled expression.
- Reconstruct observed expression with shared biological factors followed by batch-specific gene offsets and positive scales.
- Define within-batch strata from current latent factors and optimize robust worst-stratum reconstruction risk with conservative regularization.
- Output corrected expression as the shared nonnegative reconstruction and use factor vectors as the integrated representation.
- Evaluate whether shared abundant populations mix while rare or batch-restricted marker programs remain distinct.

Integration Hypothesis 6

Problem. Batch-specific and platform-specific variation can distort integrated single-cell transcriptomic embeddings and obscure shared biology, motivating invariance regularization for alignment and batch correction. In cross-cohort count data, the concrete failure mode is overcorrection: rare or cohort-enriched populations may be compressed into abundant neighboring lineages when integration optimizes global batch mixing or average reconstruction. The task is to learn a low-dimensional embedding that removes technical variation among shared cell types while preserving rare biological states and their marker-gene signal, using batch labels during training and cell-type annotations only for evaluation.

Proposed mechanism. The method is a count-based variational autoencoder with a batch-free encoder and a batch-conditioned negative-binomial decoder. Before representation learning, it estimates a simple library-size- and batch-aware null expression model, then converts cell-gene residual enrichment into marker-protection weights. These weights emphasize expression unexpectedly high within a batch, reducing emphasis on global high-abundance genes and batch-wide technical shifts. During training, the model combines the usual count reconstruction and latent regularization with a protected reconstruction distortion computed from marker-weighted negative-binomial errors. Instead of averaging this protected error over all cells, it minimizes a conditional tail-risk objective over the worst-reconstructed fraction of cells, making failures on rare-marker programs expensive even when common cell types are reconstructed well. An optional residual-program cosine reconstruction term can further protect coordinated marker patterns. The encoder mean is exported as the integrated embedding, while decoder batch conditioning absorbs nuisance technical effects.

Key steps.

- Fit a library-size- and batch-aware null model to estimate residual cell-gene enrichment.
- Convert positive residual enrichment into marker-protection weights that downweight batch-wide shifts.
- Train a batch-free encoder and batch-conditioned negative-binomial decoder on raw counts.
- Add a marker-weighted tail-risk reconstruction penalty over the worst protected reconstruction errors.
- Export the encoder mean as the integrated embedding for clustering, visualization, and scoring.

Integration Hypothesis 7

Problem. In reference-based single-cell RNA-seq annotation, a query dataset may contain cell types or states that are not present in the labeled reference. In these type-mismatched settings, standard transfer pipelines can still project every query cell onto the reference manifold and assign the nearest known label with high confidence. The concrete failure mode is unknown cell type misclassification: absent populations are over-aligned to abundant related lineages, then reported as confident false labels rather than being rejected. The task is to produce an embedding or transfer representation, calibrated per-cell confidence, and an unknown-population decision while retaining accurate labels for genuinely shared cell states.
Proposed mechanism. Model the reference and query as empirical single-cell distributions with a shared component plus a query-specific contamination component. The method estimates per-query-cell inlier weights during a robust, entropy-regularized transport alignment to the reference. Cells with plausible counterparts in the labeled reference obtain high weights and contribute strongly to label transfer; cells whose alignment would require large transport distortion receive low weights, limiting their influence on the shared manifold and preventing forced assignment to nearby known types. The final representation combines weighted cross-dataset couplings for likely shared cells with within-query neighborhood preservation for all cells, so unsupported populations remain locally coherent. Label probabilities are computed only through high-confidence couplings, while an unknown decision is derived from the inlier weight, transport cost, and neighborhood-level consistency. A speculative addition is post hoc probability calibration using held-out known classes to improve threshold stability across datasets.
Key steps.

- Normalize counts, derive a low-dimensional expression representation, and construct within-dataset neighborhoods.
- Fit a robust contaminated-distribution alignment between reference and query with learnable per-query inlier weights.
- Transfer known labels through high-weight couplings while bounding the effect of low-weight query cells.
- Embed cells using weighted cross-dataset alignment plus within-query neighborhood preservation.
- Call unknown populations from low inlier weight, high alignment distortion, and weak local label consistency.

Integration Hypothesis 8

Problem. Heterogeneous scRNA-seq datasets generated across different laboratories, platforms, tissues, species, and batches often contain technical expression biases that dominate biological variation. In such cohorts, cells may cluster by batch or time point rather than true cell identity, especially when sample composition is imbalanced and small populations are present in only a few samples. The task is to produce a batch-aware integrated representation that removes technical structure among abundant shared populations while avoiding overcorrection of rare or weakly marked cell states whose apparent sample enrichment may reflect real biology rather than artifact.
Proposed mechanism. The method treats batch integration as a local empirical Bayes partial-pooling problem in a low-dimensional expression space. Starting from normalized highly variable genes and an initial embedding, it constructs overlapping transcriptomic neighborhoods and decomposes each cell position into a neighborhood biological center, a global batch offset, a local batch residual, and noise. Global offsets are estimated across broad shared structure and removed for all cells. For each neighborhood, a batch-size-adjusted occupancy model estimates posterior cross-sample support, distinguishing broadly shared regions from sample-enriched regions. A local reliability estimate then measures whether the neighborhood-specific residual is stable enough to subtract. The final correction subtracts local residuals only in proportion to both cross-sample support and residual reliability, making correction strong in abundant shared populations and conservative where rare-cell support is uncertain. The method also returns per-cell diagnostics summarizing support and correction strength.
Key steps.

- Compute an initial low-dimensional representation from normalized, feature-filtered expression data and define overlapping local neighborhoods.
- Estimate global batch offsets and neighborhood-specific residual batch deviations in the embedding space.
- Infer batch-size-adjusted posterior support for each neighborhood using sample occupancy rather than raw mixing alone.
- Compute local empirical Bayes reliability weights for residual correction within each batch-neighborhood pair.
- Generate the integrated embedding by subtracting global offsets everywhere and shrinking local residuals according to support and reliability.

Integration Hypothesis 9

Problem. Unsupervised clustering followed by marker-based labeling remains an unreliable route to scRNA-seq cell annotation: it yields rough classifications, often estimates the wrong number of classes, and misses fine-grained similar subtypes, including closely related peripheral-blood T-cell subclasses. The task is to learn an integrated representation for multiple annotated count matrices with batch identifiers and harmonized labels that makes downstream clustering and visualization more faithful. Same-label cells should mix across batches only when support indicates the class is genuinely shared, different labels should stay separated, and rare or batch-specific annotated populations should not be absorbed into abundant related lineages.
Proposed mechanism. The method trains a neural encoder from normalized expression profiles to a low-dimensional integrated embedding using two coupled losses. First, a label-conditioned partial transport objective aligns cells only within the same harmonized label, but constrains the amount of cross-batch transported mass for each label-batch pair. These capacity limits are derived from the label-by-batch support table, assigning high alignment capacity to well-supported shared labels, partial capacity to strongly imbalanced labels, and near-zero capacity to labels that appear batch-specific. Second, a rare-balanced prototype separation objective learns one or more label prototypes and weights rare labels by effective sample size, optionally using angular margins to prevent rare populations from collapsing into abundant transcriptionally similar lineages. The output is a standard embedding for clustering, visualization, and integration assessment.
Key steps.

- Normalize count matrices, harmonize labels, and compute label-by-batch support and imbalance summaries.
- Estimate admissible same-label cross-batch alignment capacity for every label-batch pair.
- Train an expression encoder with capacity-constrained same-label partial transport in the embedding space.
- Add rare-weighted prototype separation so annotated rare labels remain distinct from nearby abundant labels.
- Export the integrated embedding and evaluate downstream clustering, visualization, and label preservation.

Integration Hypothesis 10

Problem. In multi-sample single-cell expression integration for drug-response studies, independent-cell modeling ignores intercellular interactions. Existing integration workflows often represent cells as isolated feature vectors and optimize batch mixing, so they can erase cell-cell relationship structure and tumor-microenvironment context that shape heterogeneous functional states. In compositionally imbalanced cohorts, rare malignant, stromal, or immune neighborhoods may appear in only some samples, with no valid counterpart elsewhere; forced matching then absorbs them into dominant states. The task is to learn an integrated embedding and neighbor graph that corrects technical effects among genuinely shared states while preserving cohort-restricted interaction neighborhoods relevant to response heterogeneity.
Proposed mechanism. The method represents each sample or batch as a metric-measure graph rather than a bag of independent cells. Nodes carry normalized high-variance expression or reduced components; edges connect local transcriptional neighborhoods, and graph distances summarize local cell-cell context. Pairwise batch alignment is posed as entropic unbalanced fused relational optimal transport: a transport plan is rewarded when matched cells are similar in expression and their within-batch relational distances to other cells are geometrically consistent, but the marginal constraints are relaxed so unsupported cells may remain partially unmatched. Reliable transport mass is converted into cross-batch edges only for states with both transcriptional and neighborhood support. Cells with low matched mass keep their within-batch neighborhood edges, preserving cohort-restricted tumor-microenvironment neighborhoods. The union of retained within-batch edges and confidence-weighted cross-batch edges is embedded with a graph spectral or diffusion procedure to produce a standard low-dimensional representation and neighbor graph for clustering, visualization, and response-associated heterogeneity analyses.
Key steps.

- Build a within-batch cell-neighborhood graph from normalized expression features and compute graph-derived relational distances.
- Assign cell masses and solve unbalanced fused relational transport between batches using expression similarity and graph-geometry consistency.
- Threshold or weight high-confidence transport mass to create cross-batch edges for shared states.
- Retain mostly within-batch edges for cells with low matched mass to preserve cohort-restricted neighborhoods.
- Embed the integrated graph and export the embedding, neighbor graph, and matched-mass diagnostics.

Reference mapping

Reference Mapping Hypothesis 1

Problem. Reference-based single-cell RNA-seq annotation often assumes the labeled reference and deployed query are exchangeable, but practical queries can differ by patient, disease state, tissue composition, or protocol. Thus, $P_{\text{train}}(x, y) \neq P_{\text{test}}(x, y)$, where x is the expression feature vector and y is the cell-type label. Here, P_{train} and P_{test} denote the joint distributions in the reference and query populations, respectively. Consequently, benchmark accuracy on near-i.i.d. splits can be overly optimistic about deployment reliability. In this setting, a labeled reference is used to annotate query cells in the same gene space, including known populations under donor or study shift and cells not adequately represented in the reference. The task is to assign supported known labels, reject unsupported cells as novel, and quantify per-cell uncertainty while auditing false rejection of known cells.
Proposed mechanism. The method trains a label-conditioned denoising autoencoder on reference cells after reference-defined gene selection, library-size normalization, log transformation, and reference-based gene scaling. For each candidate known label, the decoder reconstructs a query expression vector under that label, yielding a candidate-specific reconstruction residual; this tests absolute support by a reference cell-type manifold rather than relative classifier confidence. A held-out, stratified reference calibration split is excluded from fitting and model selection and supplies cell-type-specific residual distributions. Within each cell type, donor or study residual variation estimates a one-dimensional Wasserstein envelope, used to conservatively expand calibration quantiles before p-value calculation. Each query cell receives one robust conformal p-value per known label. If all p-values are at or below the chosen significance level, the cell is returned as novel/unsupported; otherwise, a lightweight classifier head ranks only labels whose p-values pass the support test. No additional biological priors or external annotations are required.
Key steps.

- Preprocess reference and query cells using transformations fixed from the reference only.
- Train a label-conditioned denoising autoencoder and ranking head on known reference cells.
- Compute candidate-specific reconstruction residuals for every query cell under every known label.
- Calibrate cell-type-specific residual p-values with donor- or study-shift-expanded conformal nulls.
- Accept supported labels, rank only accepted labels, and reject cells with no accepted label.

Reference Mapping Hypothesis 2

Problem. Across single-cell RNA-seq studies, the same biological populations are often annotated at incompatible resolutions: one study may provide only broad parent classes, while another labels fine subtypes. When these annotations are merged and treated as one flat label set, supervision becomes ambiguous because broad labels can contain multiple unobserved descendants. This harms cross-study transfer and reference mapping, especially when query cells resemble a lineage represented in the reference but not a specific known subtype. The task is to learn from mixed-depth hierarchical labels and return the deepest label supported by expression evidence rather than forcing every query cell into a fine class.
Proposed mechanism. The method models annotation as a hierarchy-consistent probabilistic descent through a cell-type taxonomy. RNA profiles from reference and query data are projected into a shared feature space using common variable genes and a neural encoder. At each internal taxonomy node, a local conditional classifier estimates probabilities for each observed child and for terminating at the current node. The probability of a fine subtype or unresolved ancestor is computed as the product of conditional decisions along its path, ensuring that child probabilities remain nested under their parents. Training uses a censored hierarchical likelihood: fully resolved cells supervise the complete path to their leaf, while cells annotated only to a broader class supervise the path only up to that class and do not impose arbitrary descendant labels. To learn behavior when fine types are absent, training episodes temporarily remove selected subtrees and train their cells to terminate at the nearest visible ancestor. Post hoc node-wise calibration supports reliable depth selection and unresolved-cell flagging.
Key steps.

- Build a label tree from all available annotation levels, allowing internal nodes to be valid outputs.
- Encode normalized shared-gene expression into a latent representation with study-aware training splits.
- Train node-local child-versus-termination probabilities using censored hierarchical likelihood for mixed-depth labels.
- Simulate missing fine subtypes by hiding subtrees during training and supervising termination at visible ancestors.
- Calibrate node-wise continuation decisions and output a path, deepest-supported label, and unresolved flag.

Reference Mapping Hypothesis 3

Problem. Non-directional integration causes over-integration in single-cell RNA-seq reference mapping. Symmetric integration or uncontrolled mapping can collapse distinct cell identities into averaged representations and can hallucinate reference-specific structure when batch compositions differ. The task is therefore to perform an explicitly directed projection of query cells into a user-chosen labeled reference while preserving the reference geometry. The method must identify query cells that are novel, poorly represented, or incompatible with the reference before label transfer, rather than forcing every cell into the nearest known label. It should provide calibrated per-cell rejection decisions and retain accurate annotation for compatible cells.
Proposed mechanism. The method fixes a labeled reference embedding and maps each query cell into that geometry without allowing the query to reshape the reference. It then tests whether each projected query cell preserves its local neighborhood relationships from expression space. A calibration bank is built by repeatedly holding out reference donors, batches, or subsets as pseudo-queries, projecting them back into the remaining reference, and recording the amount of neighborhood disruption expected for represented biology under realistic variation. For each real query cell, the algorithm forms a local neighborhood in normalized high-variance gene expression space and measures how close those neighbors remain after projection, with adjustment for local reference density. The resulting nonconformity score is compared only with calibration cells from similar mapped reference regions and density regimes to obtain a one-sided calibrated p-value. A multiple-testing procedure controls false discoveries across query cells. Labels are transferred by weighted nearest neighbors only for cells passing the compatibility test.
Key steps.

- Freeze the labeled reference representation and project query cells into it directionally.
- Construct represented calibration examples by holding out reference subsets as pseudo-queries and remapping them.
- Compute a density-normalized local neighborhood disruption score for each calibration and query cell.
- Convert query scores to stratified conformal p-values using matched calibration strata.
- Control false discoveries across query cells and transfer labels only to accepted cells.

Reference Mapping Hypothesis 4

Problem. Reference-based single-cell RNA sequencing mapping pipelines often return a best reference label without clear, interpretable reliability measures or quantitative summaries of mapping quality, such as expected precision or accuracy. This creates a concrete quality-control failure: analysts cannot easily decide whether annotations are trustworthy, whether integration decisions should be accepted, or whether unsupported labels should be rejected. The task setting is a labeled reference with expression features, donor or batch metadata, and hierarchical cell annotations, plus an unlabeled query dataset. The method must assign query cells only as specifically as the evidence supports, provide calibrated uncertainty for each assignment level, and summarize mapping quality in terms usable for annotation review.

Proposed mechanism. Train lightweight hierarchical classifiers on shared expression features from the reference, using conditional prediction heads that estimate child-label probabilities only within an accepted parent category. Multiple reference-specific models are fit across donor, batch, or study partitions so that disagreement becomes an explicit uncertainty feature rather than being hidden by a single pooled predictor. For each query cell at each hierarchy node, collect local evidence: mean child probabilities, entropy, vote disagreement, sibling probability margins, and total probability mass assigned to plausible descendants. A calibration layer is learned from held-out reference cells and leave-one-child-out pseudo-novel episodes, converting these local evidence features into calibrated continuation risks for choosing each child. Inference follows a risk-limited stopping rule: descend only when the calibrated risk is below a prespecified resolution value; otherwise stop at the deepest supported parent and mark lower levels as abstained. Outputs include accepted labels, abstention flags, calibrated risks, and dataset-level quality summaries.

Key steps.

- Align reference and query to shared expression features and organize labels as a parent-child taxonomy.
- Train partition-aware hierarchical classifiers with conditional child-label heads under each parent node.
- Calibrate continuation risk using held-out shared cells and pseudo-novel child-withholding episodes.
- Traverse the taxonomy for each query cell, descending only when calibrated risk supports a finer label.
- Report per-cell accepted labels, abstentions, uncertainty scores, and aggregate mapping-quality summaries.

Reference Mapping Hypothesis 5

Problem. The problem is labor-intensive subjective manual cell labeling in single-cell RNA sequencing analysis. The standard workflow of unsupervised clustering followed by marker-based manual labeling is time-consuming, subjective, and especially difficult when the number of cell types is unknown. In a practical task setting, a labeled reference expression dataset with counts, cell-type labels, and batch metadata is available, and a query dataset has matched genes. The goal is to assign each query cell or metacell to a supported reference cell type, or abstain as unassigned or novel, while avoiding confident transfer of the closest known label when the query contains unsupported cell types or misleading nuisance programs.
Proposed mechanism. The method treats annotation as controlled discovery of biologically compatible query-to-reference label edges rather than forced classification. For each reference cell type, it learns a compatibility contract containing identity genes, essential markers, expected marker-expression ranges, and excluded nuisance genes. These genes are selected using differential expression, specificity, and stability across batches, while removing signals associated with batch, stress, cell-cycle, mitochondrial, ribosomal, or broadly expressed programs. For every query cell or metacell and every reference type, the method computes a compatibility statistic that combines proximity to that type over identity genes with an aggregate penalty for missing essential markers. The statistic is calibrated against held-out non-target reference cells to obtain an empirical compatibility p-value for each query-type edge. A multiple-testing procedure controls the false discovery rate across the full query-by-type graph. A query receives a label only if one compatible edge is discovered, is unassigned if none are discovered, and is ambiguous if multiple compatible edges remain.
Key steps.

- Learn batch-stable identity genes, essential markers, marker expectations, and nuisance exclusions for each labeled reference cell type.
- Represent each query cell or metacell using only retained identity and marker evidence matched to reference genes.
- Score every query-type edge by combining identity-gene proximity with an aggregated deficit of essential-marker support.
- Calibrate edge scores using held-out non-target reference cells to produce empirical compatibility p-values.
- Apply graph-wide false discovery rate control and assign unique, unassigned, or ambiguous outcomes.

Reference Mapping Hypothesis 6

Problem. Standard scRNA-seq annotation often relies on unsupervised clustering followed by manual marker inspection, which is labor-intensive and produces labels that vary across analysts, references, preprocessing choices, and datasets. An automated pipeline is needed for assigning unlabeled query cells using a labeled reference while avoiding the closed-set failure mode in which every query cell is forced into a known reference type. The task is to output known labels where reference support is sufficient, explicit unassigned calls where cells are likely unsupported by the reference, calibrated per-cell support and novelty evidence, and a query-level estimate of the unsupported-cell fraction with uncertainty.

Proposed mechanism. The method is a post-processing layer for reference-based scRNA-seq mappers that converts their raw label scores, similarities, margins, entropies, probabilities, or nonconformity scores into calibrated reference-support scores. Calibration is performed within each predicted label by comparing a query cell's uncertainty profile with held-out reference cells assigned to the same label, reducing dependence on score scale, cell-type abundance, and label granularity. To learn what unsupported cells look like, annotated reference populations are repeatedly removed at cell-type, subtype, or state resolution, remapped against the incomplete reference, and used as pseudo-novel positive examples. The real query score distribution is then treated as an unlabeled mixture of reference-supported and reference-unsupported cells. A positive-unlabeled mixture proportion estimator finds the largest pseudo-novel component compatible with a valid residual known-cell distribution. Cells are ranked by posterior novelty evidence, the estimated unsupported fraction is assigned as unassigned, and bootstrap resampling provides uncertainty intervals.

Key steps.

- Map each query cell to the labeled reference and collect raw label scores and uncertainty summaries.
- Transform raw scores into label-conditional empirical reference-support scores using held-out known reference cells.
- Generate pseudo-novel score distributions by leave-population-out remapping across annotation granularities.
- Estimate the query-specific unsupported-cell fraction through positive-unlabeled mixture proportion estimation.
- Assign unassigned calls to the highest-novelty cells while retaining known labels for the remaining cells.

Reference Mapping Hypothesis 7

Problem. Million-cell single-cell RNA-seq atlases create compute and memory bottlenecks because full expression matrices, neighbor structures, repeated projection steps, visualization, differential analyses, and model training all require expensive passes over many redundant cells. The concrete failure mode is that common populations dominate storage and minibatches, while naive subsampling reduces cost but can erase rare populations, boundary states, and local expression programs needed for downstream validity. The task is to construct a representative sketch that is much smaller than the original atlas, yet preserves major cell-type geometry, local heterogeneity, rare-state support, and compatibility with downstream visualization, analysis, and machine-learning workflows.
Proposed mechanism. The method compresses an atlas into calibrated representative metacells rather than individual cells. It first forms expression metacells and records their size, local density, and annotation or cluster stratum. Within each stratum, it constructs local convex hulls from nearby metacells and uses simplex-constrained least-squares projection to measure how well omitted metacells can be reconstructed from candidate representatives. A metacell is retained when its removal increases the stratum-specific projection distance, the gap between stratum-constrained and all-atlas projections, or a coherent residual expression score. Held-out metacells calibrate acceptable reconstruction thresholds by stratum, density, and metacell size, preventing dense populations from consuming the sketch while sparse populations are overcompressed. The final sketch stores representatives, weights, calibrated reconstruction bounds, and residual summaries. A speculative addition is to expose the weighted representatives directly as training units for neural models; this is not established by the projection evidence and must be evaluated separately.
Key steps.

- Aggregate cells into metacells and annotate each with size, local density, and biological or data-driven stratum.
- Build local stratum-specific convex hulls and compute constrained projection distances for candidate representative sets.
- Retain metacells that support hull geometry, reduce global-versus-stratum projection gaps, or explain coherent residual programs.
- Calibrate reconstruction and residual thresholds on held-out metacells within density- and size-aware strata.
- Export the compact weighted sketch with compatibility diagnostics for visualization, analysis, and model-training experiments.

Reference Mapping Hypothesis 8

Problem. When integrating many GB cohorts, batch-correction can fail in two opposing ways: over-correction may collapse biologically meaningful cell-state structure, while under-correction may leave cohort-specific technical variance that masquerades as state differences. The task is to support benchmark-driven selection of an integration approach by measuring whether aligned cells are explainable as known cellular baselines plus cohort-shared perturbation, or instead retain residual expression inconsistent with any represented state. The setting includes labeled or partially labeled reference cohorts, unlabeled target cohorts over shared genes, heterogeneous cohort effects, and possible cell states absent or weakly represented in the reference.
Proposed mechanism. The method builds an integration-selection score around perturbation-quotient residuals rather than relying only on visual mixing or neighbor agreement. For each annotated cell category, it fits a low-rank affine baseline model capturing within-category biological variation, while preserving discriminative marker directions as protected identity axes. After provisional mapping of target cells to candidate baselines, it estimates cohort-shared perturbation factors from residual directions that recur across multiple provisional categories and are weakly aligned with protected identity axes. Each candidate integration is then evaluated by projecting out these allowable shared perturbation factors and measuring the remaining label-specific residual. Held-out cohort or donor splits provide calibration of residual distributions, producing category-wise support values, ambiguous sets, and unsupported calls. Over-correction is indicated when protected identity axes or calibrated residual structure are inappropriately removed; under-correction is indicated when recurrent technical residuals remain after allowable perturbation removal. The speculative addition is a multi-objective selector that ranks integrations by combining calibrated support, residual preservation, and cohort-mixing diagnostics.
Key steps.

- Fit category-specific low-rank baseline subspaces and protected identity directions from labeled reference cohorts.
- Provisionally assign target cells to candidate baselines and compute category-specific residuals.
- Estimate recurrent cohort-shared perturbation factors while penalizing alignment with protected identity directions.
- Compute perturbation-quotient residuals and calibrate category-wise support using held-out cohort or donor splits.
- Rank candidate integrations by calibrated support, residual novelty, identity preservation, and remaining cohort effect.

Reference Mapping Hypothesis 9

Problem. Existing external atlas-mapping tools are often inseparable from the integration framework that created the atlas, requiring original latent transforms, software objects, or raw reference data that may be unavailable after publication. This prevents reliable reuse of published single-cell RNA atlases as fixed coordinate systems and makes cross-study comparisons difficult to reproduce when the original pipeline cannot be rerun. The task is to map new query cells onto an annotated reference atlas using only matched reference expression and a fixed published low-dimensional embedding, while reporting projected coordinates, optional labels, and calibrated per-cell coordinate reliability so weakly supported or atlas-missing states are not displayed as falsely precise points.
Proposed mechanism. The method represents the published atlas by selected cells or landmarks in its fixed embedding and learns an expression-to-coordinate compatibility function without needing the original integration transform. For each reference cell, a lightweight supervised model is trained to assign low energy to nearby landmarks and higher energy to distant landmarks, using soft targets shaped by neighborhood topology and locally scaled embedding distance. For a query cell, the energy landscape yields a point projection by minimum-energy or energy-weighted averaging, but the primary output is a calibrated coordinate envelope: all atlas locations whose energies remain statistically compatible with the query profile. Split calibration on held-out reference cells converts raw energies into nominal coverage thresholds, stratified by local density or geometry to avoid overly broad global rules. Additional leave-state-out calibration with annotated groups withheld mimics missing biological support, teaching envelopes to expand when a query resembles a lineage neighborhood but lacks a well-supported atlas location.
Key steps.

- Select representative atlas landmarks and compute local topology, density, and embedding-distance summaries.
- Train an expression-to-landmark energy model using topology-aware soft coordinate targets from reference cells.
- Calibrate energy thresholds on held-out reference cells within local geometry or density strata.
- Repeat calibration with annotated groups withheld to estimate envelope inflation for weakly covered states.
- For each query cell, output a projected coordinate, optional label summary, coordinate envelope, and reliability scores.

Reference Mapping Hypothesis 10

Problem. Marker-driven or reference-derived annotation workflows often rely on a small set of canonical genes or on global transcriptomic similarity, but they do not effectively exploit the much larger pool of informative non-marker genes. In spatially continuous tissues, this creates a concrete failure mode: cells or spots that satisfy familiar marker expectations are assigned confidently even when broader supportive expression is weak, contradictory, or spatially inconsistent. Conversely, noisy or missing genes may be treated as negative evidence rather than uncertainty. The task is to annotate spatial or spatially registered transcriptomic observations using shared genes from a labeled reference while producing calibrated confidence and interpretable evidence from markers, non-marker genes, contradictions, and ignorance.
Proposed mechanism. The method would learn, for each reference cell type, a type-specific partition of genes into essential marker genes, supportive non-marker genes, lateral contradictory genes, and noisy or excluded genes, using expression effect size, detection stability, and reproducibility across reference subsets. For each query observation and candidate label, genes are converted into evidential scores representing support, contradiction, or ignorance rather than a single similarity value. Reliably absent essential markers contribute contradiction, while dropout-prone or weakly measured genes mainly contribute ignorance. Supportive non-marker agreement is accumulated separately from marker evidence so that marker-positive but broadly unsupported observations can be flagged. A cautious evidence-combination rule preserves conflict instead of normalizing it away. The resulting vector of support, contradiction, conflict, ignorance, marker support, non-marker agreement, and neighborhood agreement is calibrated using held-out reference splits and pseudo-novel splits. The neighborhood agreement term is a speculative addition, intended to encode spatial continuity without overriding cell-intrinsic evidence.
Key steps.

- Partition genes per reference label into marker, supportive non-marker, contradictory, and unreliable sets.
- Convert query expression for each label into support, contradiction, and ignorance masses at gene level.
- Fuse gene-level masses while preserving conflict between marker support and broader evidence.
- Add a calibrated spatial neighborhood consistency score as an explicit, down-weightable evidence source.
- Output a known label or unassigned call with calibrated confidence and evidence-based explanation.

Supplementary Note 5: Key prompts and decision boundaries

This note reports the key instructions used to construct RPKGs, reformulate Problems and Mechanisms, formulate candidate method hypotheses, and assess selected outputs. The Requirement and Articulation instructions are reproduced in full; the remaining boxes give the operative criteria and output formats, with formatting wrappers omitted. In the prompts, "Design" denotes the Mechanism primitive, the field required mechanism type holds the Requirement, and "batch" refers to the set of records submitted in one model call. Decisions about whether to pursue a hypothesis remained with the researchers.

RPKG construction

RPKG schematization and relation extraction

ROLE

You are a Schematization Agent for single-cell genomics research. From one paper's primitive-aware distilled summary you build that paper's slice of the Research Primitive Knowledge Graph (RPKG): first extract the typed primitives (nodes), then extract the typed relations between them using only the primitives you extracted, and return one consolidated JSON object.

PROTOCOL

1. Execute the primitive-extraction instructions first.
2. Use only the primitives you extracted in this same response when applying the relation-extraction instructions.
3. Return only the final consolidated JSON object specified at the end.
4. Do not emit intermediate arrays, explanations, or markdown fences.

METHOD primitives

Extract named computational methods, software tools, and analysis pipelines mentioned in the summary. Only standalone, named tools or algorithms with their own identity that could be cited or downloaded independently (e.g., Harmony, scVI, Seurat, Combat, LIGER, MNN and Scanorama). Do not extract generic operations or building blocks (PCA, k-means, SVD, t-SNE or UMAP) unless they are the paper's primary subject or novel contribution; internal sub-steps of a larger method; broad categories; or standard library utilities. For each Method, return its name, type, language, URL, limitations, novelty and whether it was proposed in the focal paper. Use only explicitly named methods; do not infer.

PROBLEM primitives

A ProblemNode is a specific technical or biological problem that motivates a mechanism in this paper. Extract only the primary problems that the paper directly addresses, typically three to eight. If two problems are paraphrases or facets of one challenge, consolidate them. For each Problem, return a concise name, one precise paper-specific description and a problem type. Do not extract general domain challenges that are merely acknowledged, discussion-only limitations, duplicate restatements, vague challenges without a mechanism link, or limitations of baselines unless they are used as motivation.

MECHANISM primitives

A MechanismNode is a core technical choice that defines the proposed method's algorithmic strategy. Extract only the primary mechanisms, typically three to eight. If two descriptions are sub-aspects of the same architectural choice, consolidate them under the broader choice. For each Mechanism, return a concise name, a one-sentence description of what it does and the specific algorithmic or architectural technique. Do not extract routine preprocessing, hyperparameter or optimizer details, generic components, baseline mechanisms, or duplicate restatements. A retained mechanism must be explicitly part of the proposed method and must distinguish it from alternatives.

EFFECT and TASK primitives

An EffectNode is a concrete outcome or improvement caused by a mechanism. A TaskNode is a downstream biological or analytical task that the method enables or materially improves. Extract approximately three to six Effects and two to five Tasks, consolidating duplicates. Effects must be explicit computational or methodological outcomes of a mechanism rather than intermediate computations, isolated benchmark scores or biological findings. Tasks must be explicit downstream goals rather than internal processing steps. Standardize synonymous task names.

RELATIONS

Link the extracted primitives with typed relations, using primitive names exactly as extracted. Relations may be implicit and should be inferred from meaning rather than keyword matching.

- `built_on` (MethodNode → MethodNode): builds upon or extends a foundational method.
- `compared_to` (MethodNode → MethodNode): benchmarked against, with comparison equal to outperforms, underperforms, mixed or unknown.
- `contains` (MethodNode → MethodNode): a pipeline contains an algorithm component.
- `published_in` (MethodNode → PaperNode): the method is introduced in this paper.
- `motivates` (ProblemNode → MechanismNode): a problem led to a mechanism.
- `achieves` (MechanismNode → EffectNode): a mechanism produces an effect.
- `enables` (EffectNode → TaskNode): an effect enables or improves a downstream task.

- `addresses` (MethodNode → ProblemNode): a method targets a problem.
- `incorporates` (MethodNode → MechanismNode): a method is built around a mechanism.
- `limits` (ProblemNode → MethodNode): an existing method has this limitation.

For `built_on`, the source is the newer or dependent method. For `compared_to`, the source is the focal method and the target is the baseline. For `contains`, the source is the pipeline. For `published_in`, the source is the method. Prefer a defensible relation over omission, require source and target types to match the relation schema, and return an empty relation array when none is supported.

CONSOLIDATED OUTPUT

Return one JSON object and no markdown fence:

```
{
  "pmc_id": "...",
  "paper_title": "...",
  "method_nodes": [],
  "problem_nodes": [],
  "mechanism_nodes": [],
  "effect_nodes": [],
  "task_nodes": [],
  "relations": [
    {
      "source_node": "...",
      "target_node": "...",
      "relation_type": "...",
      "source_node_type": "...",
      "target_node_type": "...",
      "comparison": ""
    }
  ]
}
```

Internal hypothesis revision

The following prompts were used by the Critic in the topic-specific workflow. Labels such as "PROCEED" and "pass" controlled routing between revision rounds only.

Critic: direction review

You are a senior bioinformatics researcher evaluating whether a research direction is worth pursuing.
You have been provided with (1) the proposed motivation and computational problem, (2) search results showing existing work in this area and (3) the original research topic or direction. The proposed method hypothesis must stay within the original direction. If the motivation or computational problem drifts substantially into an adjacent task or modality, assign `ABANDON` regardless of the other dimensions.
The problem must be a real, recurring practitioner pain point rather than a niche theoretical concern. Evaluate the following dimensions on a 1–10 scale.

DIRECTION SATURATION

Despite the legacy name, this axis evaluates the direction rather than the number of tools in the broader area. Score three properties together: (a) whether the mechanism causally addresses the problem; (b) whether the direction stays within the supplied topic; and (c) whether any retrieved work already uses the same mechanism for the same problem framing.

- 9–10: all three properties hold; no retrieved work uses the same mechanism for the same problem framing.
- 7–8: coherence and topic fit hold; at most one minor overlap uses a related mechanism for a different variant or a different mechanism for the same problem.
- 5–6: the mechanism is coherent but the proposal shows mild scope drift, or two to three works show partial mechanism–problem overlap.
- 3–4: substantial scope drift, or at least three works directly match the mechanism–problem combination.
- 1–2: the causal pathway fails, the proposal is fundamentally off-topic, or a retrieved work is a semantic clone.

Do not lower this score merely because the broader topic is crowded.

DIFFERENTIATION SPACE

- 9–10: the angle is fundamentally different from all existing work.
- 7–8: clear technical differentiation from the closest work.
- 5–6: some differentiation but substantial overlap.
- 3–4: a minor variation of existing work.
- 1–2: essentially the same as existing work.

IMPACT POTENTIAL

- 9–10: would transform how the field approaches the problem.
- 7–8: a substantial practical improvement for real workflows.
- 5–6: useful but modest improvement.
- 3–4: marginal benefit over existing solutions.
- 1–2: no meaningful practical impact.

EVALUABILITY

- 9–10: established open benchmarks and standard metrics are available.
- 7–8: suitable open data and metrics are available, with runnable baselines.
- 5–6: data require curation or metrics are incompletely standardized.
- 3–4: no established benchmark; data require special access or substantial manual curation.
- 1–2: the proposal cannot be evaluated with currently available public data.

STORY CLARITY

- 9–10: one surprising, testable claim demands a new method and implies a clear key figure.
- 7–8: the claim is clear and testable but could be more distinctive.
- 5–6: a problem is described without a single testable claim.
- 3–4: the direction is a vague request for a better tool.
- 1–2: there is no discernible claim or story.

PUBLISHABILITY

- 9–10: the proposal targets an important under-served variant of a recognized problem; existing baselines remain applicable; and the claim is mechanistic and testable across at least three datasets.
- 7–8: the proposal addresses a practically relevant problem through either a variant-specific but incremental mechanism or a crisp mechanism for the standard problem.
- 5–6: the proposal addresses a known problem under standard assumptions, relies mainly on technique-level improvement, or has a thin benchmark plan.
- 3–4: the proposal is technique-driven without biological motivation, or the problem is too isolated to support comparison.
- 1–2: the problem is artificial, already solved or unlikely to interest practitioners.

Explicitly test the variant-angle pattern: identify the well-studied problem, the relaxed assumption and the existing baselines that remain runnable on the harder case.
Use the internal routing labels `PROCEED` when every dimension is at least 8; `PIVOT` when impact or differentiation is at least 7 but another dimension is below 8; and `ABANDON` when direction saturation or impact is at most 3, evaluability at most 4, story clarity at

most 3, or publishability at most 3.
Return one JSON object:

```
{
  "scores": {
    "direction_saturation": N,
    "differentiation_space": N,
    "impact_potential": N,
    "evaluability": N,
    "story_clarity": N,
    "publishability": N
  },
  "similar_works": ["title - overlap description"],
  "feedback": "detailed evidence-linked feedback",
  "story_assessment": "testability and distinctiveness of the claim",
  "publishability_assessment": "scope, relevance and methodological basis",
  "verdict": "PROCEED|PIVOT|ABANDON"
}
```

Mechanism review 1/4: biology specialist

You are a biology-grounding reviewer for a bioinformatics method hypothesis. Evaluate only whether the method addresses a real biological problem with sound biological motivation. The strongest hypotheses solve a problem that practitioners encounter rather than a purely computational curiosity.

RUBRIC (1–10)

- 9–10: identifies topic-specific biological challenges and addresses them through explicit mechanisms; the challenge–component mapping is clear.
- 7–8: identifies real biological challenges and addresses most of them, with only minor mapping gaps.
- 5–6: mentions biological challenges, but the method is generic and lacks corresponding mechanisms.
- 3–4: uses biological framing as an afterthought to a primarily technical proposal.
- 1–2: identifies no real biological problem.

WHAT TO CHECK

Determine whether the proposal identifies concrete, topic-specific biological challenges rather than vague difficulties. Check whether each challenge is addressed by a specific component, whether the biological claims are accurate, whether known biological constraints are respected, whether relevant information is preserved, and whether a practising biologist would benefit from the tool. A proposal that acknowledges biological challenges but does not act on them should score at most 5.

The internal pass threshold is 8. Return one JSON object:

```
{
  "dimension": "biology_grounding",
  "score": N,
  "passed": true|false,
  "issue": "specific problem if any",
  "recommendation": "concrete revision if needed"
}
```

Mechanism review 2/4: novelty specialist

You are a technical-novelty reviewer for a bioinformatics method hypothesis. Evaluate only whether the core contribution is genuinely new rather than an assembly of established techniques. Prefer one clear technical insight over several stacked innovations.
You receive literature and RPKG search results and, when available, prior rejected candidates and RPKG baseline methods. Classify the candidate before scoring it; the class determines the score ceiling.

NOVELTY CLASSES

- **Class A: new mathematical formalism** (ceiling 10; floor 7). A mathematical primitive from another field is central to the target problem. The proposal must name the formalism, source field and prior application.
- **Class B: new invariance, equivariance or symmetry constraint** (ceiling 9; floor 7). The architecture or objective enforces a new structural property, expressible as $f(g \cdot x) = f(x)$ or $f(g \cdot x) = g \cdot f(x)$.
- **Class C: new reduction or reformulation** (ceiling 9; floor 6). The proposal reframes the problem such that existing methods become explicit special cases or a previously disconnected method class applies.
- **Class D: combination of existing mechanisms** (ceiling 6; floor 1). The novelty lies in a recipe rather than a new formalism, structural property or reformulation.

MEANINGFULNESS GATE

For Classes A–C, mentally remove the claimed formalism, constraint or reformulation. If the central claim survives, the element is decorative and the proposal must be downgraded to Class D. If removing it destroys the central claim, then it is load-bearing. For Class D, evaluate whether components form a causal chain in which one component's output controls the next, or whether they are merely independent additions. Unconnected stacking scores 1–4; a clear mechanistic chain may score 5–6 but remains subject to the Class D ceiling.
Before scoring, probe recent prior art for papers that combine the same mechanisms. Name concrete overlaps. If the core mechanism matches a prior rejected candidate, score at most 4. If it matches an RPKG baseline, score at most 5. For each comparison, state one mechanism-level difference rather than a superficial implementation change. In ambiguous cases choose the lower-ceiling class.
The internal pass threshold is 8. For a Class D candidate, recommend one concrete Class A, B or C reformulation for the next revision.
Return one JSON object:

```
{
  "dimension": "technical_novelty",
  "score": N,
  "passed": true|false,
  "issue": "specific overlap, failed meaningfulness test, or prior art",
  "recommendation": "mechanism-level revision",
  "distinguishing_mechanism": "difference from the closest precedent"
}
```

Mechanism review 3/4: feasibility specialist

You are a feasibility and scope reviewer for a bioinformatics method hypothesis. Evaluate only whether the method can be built and tested within a focused project. Strong methods typically contain two or three core components rather than a large collection of modules or objectives.

RUBRIC (1–10)

- 9–10: one or two core components, a clear implementation path and fewer than ten ablation experiments.
- 7–8: two or three components, manageable scope and tractable ablation.
- 5–6: at least four components or loss terms, with a large ablation matrix or weakly justified parts.
- 3–4: a kitchen-sink architecture whose engineering or ablation burden is impractical.
- 1–2: the claimed scope substantially exceeds what the proposed study can deliver.

Check whether each module is necessary, whether the central mechanism can be explained in one sentence, how many distinct components and objectives are present, whether framing matches the intended scale, whether one researcher could implement it within three to six months, and which parts could be removed. If there are more than four loss terms or more than four architectural components, score at most 5 and name the components to remove. The internal pass threshold is 8.

Return one JSON object:

```
{
  "dimension": "feasibility_scope",
  "score": N,
  "passed": true|false,
  "issue": "source of infeasibility or excess scope",
  "recommendation": "components to remove or simplify"
}
```

Mechanism review 4/4: evaluability specialist

You are an evaluability reviewer for a bioinformatics method hypothesis. Evaluate only whether the method can be rigorously evaluated with accessible benchmarks, metrics and baselines.

RUBRIC (1–10)

- 9–10: established benchmarks, directly applicable metrics, accessible baselines and a well-defined protocol.
- 7–8: suitable benchmarks and metrics are available with minor adaptation.
- 5–6: datasets require curation, metrics are not fully standardized or some baselines lack public code.
- 3–4: no established benchmark exists for the task and a new protocol would be required.
- 1–2: meaningful evaluation is not possible with available data and metrics.

PIPELINE COMPATIBILITY

The method must consume `.h5ad` data containing matrices and cell or feature metadata, and return standard embeddings, labels or per-cell scores in an `.h5ad` object. Raw FASTQ, BAM, fragment, genome, annotation-track or external-database dependencies require an explicit pipeline extension. The execution environment is Python with standard single-cell, machine-learning and metric libraries; protocol transformations may create splits, hold-outs, modality drops, masks and deterministic subsamples. They cannot fetch new external assets or run unsupported upstream assays. If the proposal violates this contract, score at most 5, name the mismatch and suggest a matrix-level or output-compatible reformulation.

Require real biological data, scenario diversity, multidimensional metrics, published task-relevant benchmarks, installable baselines, accessible datasets, and measurements aligned with every claimed advantage. The internal pass threshold is 8.

Return one JSON object:

```
{
  "dimension": "evaluability",
  "score": N,
  "passed": true|false,
  "issue": "evaluation or execution mismatch",
  "recommendation": "a feasible validation revision"
}
```

Critic: claim-aligned benchmark review

You are a senior reviewer performing a quick sanity check on a method hypothesis's evaluation direction. This is not a detailed experiment-plan review. Check whether the proposed high-level validation is aligned with the topic and the hypothesis's claims. If a benchmark, baseline or metric belongs to a different problem, cap the affected dimension at 4.

RUBRIC (1–10)

- **Benchmark appropriateness**: whether named datasets are recognized for the problem and directly test the claims.
- **Baseline fairness**: whether four to six relevant baselines cover the major methodological categories, including current strong methods.
- **Metrics completeness**: whether the metrics directly measure every claimed improvement.
- **Plan coherence**: whether benchmarks, baselines, metrics and expected outcomes define what would count as support or failure.
- **Unique advantage**: whether one testable dimension follows from a structural property of the proposal and is not already the optimization target of a baseline.

Scores of 9–10 denote direct, comprehensive alignment; 7–8 denote good coverage with minor gaps; 5–6 denote partial alignment; 3–4 denote a substantial mismatch; and 1–2 denote an absent or contradictory plan. Do not penalize a simplified outline for lacking download links, preprocessing commands or hyperparameters.

Use `SOLID` when all five dimensions are at least 8; `NEEDS_ADJUSTMENT` when at least one dimension is 5–7; and `INSUFFICIENT` when any dimension is at most 4. These are internal revision labels only.

Return one JSON object:

```
{
  "scores": {
    "benchmark_appropriateness": N,
    "baseline_fairness": N,
    "metrics_completeness": N,
    "plan_coherence": N,
    "unique_advantage": N
  },
  "feedback": "specific claim-aligned revisions",
  "unique_advantage_assessment": "what is unique and how it could be tested",
  "verdict": "SOLID|NEEDS_ADJUSTMENT|INSUFFICIENT"
}
```

Requirement and Articulation reformulation

Problem-to-Requirement reformulation

```
You are a computational biology expert characterizing what kind of solution a research problem
    requires.

You will receive a JSON object with ‘query_papers‘. Each paper contains one or more complete ‘
    problem_records‘. Each record describes a ProblemNode from a knowledge graph of single-cell
    genomics methods, with its KG context: Designs the Problem motivates, Methods it limits,
    Methods that address it, and paper excerpts.

For the temporal experiment, every contextual neighbour and paper excerpt was published no later
     than ‘context_cutoff_year‘. The query paper contributes only its complete Problem primitive
    ; its own title, summary, novelty, Design, Method, Effect, and relations are not supplied.
    Empty context lists are valid. No similar-Problem or nearest-neighbour retrieval is
    performed.

Your task is to produce one clear ‘required_mechanism_type‘ for every supplied Problem. Despite
    the legacy field name, its content must be a **solution-agnostic capability requirement**:
    what a successful solution must accomplish and preserve, not which algorithm, architecture,
    or mathematical construction it should use.

Judge every ‘problem_record‘ independently. Use only that record’s own ‘problem‘ and its own
    eligible pre-cutoff context. Other papers or other Problem records included in the same
    batch are separate queries; never transfer a method, operation, assumption, or terminology
    from one record to another. Historical neighbours are evidence about the failure,
    constraints, and desired outcome. They are not templates whose implementation should be
    copied into the requirement.

If the current Problem names an existing method, operation, or computational bottleneck as the
    source of failure, use it only to identify the capability or constraint that a successor
    must improve. Do not assume the successor must retain that method’s algorithm family. Retain
     a named construction only when the Problem explicitly requires it as an input, output, or
    interface rather than merely reporting that it failed.

This is requirement elicitation, not solution proposal. For a runtime, memory, or scale
    bottleneck, state that the underlying scientific function must remain feasible under the
    reported resource or scale constraint while preserving the required output fidelity. Do not
    invent, select, or repeat a computational shortcut for achieving that improvement. More
    generally, every implementation-level noun phrase in the answer must be required by the
    current Problem as an invariant input, output, or interface; otherwise remove it and state
    the capability it was intended to provide.

Treat every method name, software name, acronym, and citation in the Problem as an opaque label.
     Do not expand it or use outside knowledge, model memory, or an inferred meaning of its name
     to reconstruct how that method works. Only functions, failures, inputs, outputs, and
    constraints explicitly stated in the current Problem record or its eligible context may
    appear in the requirement.

Silently perform a final what-versus-how rewrite before returning: remove any phrase that says
    how a solution would be implemented and replace it with the observable function, constraint,
     or outcome that phrase was meant to provide. Output only the rewritten requirement.

Use this abstract form as guidance, without copying the bracketed placeholders into the answer:
    "A mechanism that [performs the operation required by this Problem] while [preserving or
    remaining robust to the constraint stated by this Problem], so that [the intended effect
    stated by this Problem]."

Avoid descriptions that are merely a generic method label, merely an improvement adjective, or
    an instruction to use a particular named technique. Those formulations are respectively too
    vague, incomplete, or too implementation-specific.

The description must clearly state the functional operation or transformation that is needed,
    the property it must preserve or robustness it must possess, and the intended effect.
    Describe these at the capability level and do not name a particular implementation that
    could perform the operation.
```

```
Do not introduce a named method or software package. Do not prescribe an algorithm family, model
     architecture, objective function, representation, optimization procedure, or particular
    mathematical decomposition unless the **current Problem record itself** makes that
    construction an explicit non-negotiable input, output, or constraint. Merely seeing it in a
    historical neighbour or another batch record is not sufficient. When several different
    mechanisms could satisfy the Problem, express the shared capability and outcome constraints
    rather than choosing one implementation.

Ground the requirement in the current Problem evidence: its eligible limits and motivates
    context may establish what the solution must work around or produce, but not a preferred
    implementation. Keep it bounded to single-cell RNA analysis and do not generalize beyond
    that topic. Before returning, apply a substitution check: multiple scientifically distinct
    implementations should be able to satisfy the wording. If the wording unnecessarily selects
    one implementation family, rewrite it at the capability level.

Write as much as needed to make the operation and required property clear, without filler. This
    is prompt-level guidance, not a mechanical word-count rule. Never shorten or omit a query to
     satisfy a batch target.

Return strict JSON with this exact top-level shape:

{
  "papers": [
    {
      "source_pmc_id": "<exact supplied ID>",
      "requirements": [
        {
          "query_id": "<exact supplied ID>",
          "required_mechanism_type": "<clear abstract mechanism type>"
        }
      ]
    }
  ]
}

Return every supplied paper and every supplied query ID exactly once. Output JSON only, with no
    markdown fence or preamble.
```

Mechanism-to-Articulation reformulation

```
You are a computational biology expert distilling a mechanism from primary literature.

You will receive a JSON object with ‘source_papers‘. Each paper contains its complete
novelty summary and one record for every paper-local Design. Each Design record contains
every Effect connected to that Design by a source-local ‘achieves‘ relation. Write exactly
one mechanism articulation for each Design record: explain WHY the Design produces all of
its documented Effects.

Frame the mechanism description so it is interpretable in the context of single-cell RNA
analysis, but keep the description abstract enough that the mechanism could in principle
be transferred to other problems within the same topic.

A mechanism description should:

- State the shared causal pathway once: what the Design does -> what intermediate state it
  creates -> how every supplied Effect emerges.
- Return one coherent paragraph. Use as many sentences as required to cover every supplied
  Effect, without repeating the same Design operation or adding filler.
- Use concrete biology or computation vocabulary, not generic ML phrases like "improves
  performance".
- Be defensible from the supplied complete source-local record and novelty summary. Do not
  introduce mechanisms the evidence does not support.
- Be transferable in spirit: the core principle should be statable independently of the
  specific Design or Effect names, so the mechanism can be matched to other problems
  requiring a similar principle.

Never omit an Effect, shorten an input record, or collapse distinct Effects to satisfy a
batch target. There is no Effect-count cap.

The temporal execution changes only record availability: each articulation retains its
source publication year and a fold may later use it only when that year is no later than
the cutoff. It does not change the mechanism-distillation task above.

Return strict JSON with this exact top-level shape:

{
  "papers": [
    {
      "source_pmc_id": "<exact supplied ID>",
      "articulations": [
        {
          "articulation_id": "<exact supplied ID>",
          "mechanism_articulation": "<one complete Design-level causal mechanism covering every
    supplied Effect>",
          "covered_effect_names": ["<every supplied Effect name, copied exactly once>"]
        }
      ]
    }
  ]
}

Return every supplied paper and every supplied articulation ID exactly once. Output JSON
only, with no markdown fence or preamble.
```

MCG candidate gating, build sketches and ranking

Causal-fit gate

ROLE AND INPUT

For every candidate Problem–Mechanism pair, judge whether the Mechanism would intervene on the causal driver of the Problem. The runtime item contains the Problem-decomposition description and required mechanism type together with the candidate Design, its linked Effect and a causal description of how the Design produces that Effect. Judge the causal relation, not lexical or topical similarity.

LABELS

Assign exactly one label to every pair:

- `INTERVENTIONAL_FIT`: implementing the Mechanism would directly attack the causal pathway responsible for the Problem.
- `TOPICAL_ONLY`: the texts share a domain, vocabulary, outcome or architecture family, but the Mechanism does not act on the Problem's causal driver.
- `DIAGNOSTIC_OR_EVAL`: the candidate measures, benchmarks, simulates, profiles or reports the Problem rather than intervening on it.

Adjacent-modality methods, same-outcome descriptions with different causal paths and generic architecture matches must not be treated as interventional fits. Only `INTERVENTIONAL_FIT` items remain in the internal candidate queue.

OUTPUT

Return strict JSON with one entry for every supplied `pair_id`: `{"judgments": [{"pair_id": "...", "fit": "...", "reason": "one concrete sentence"}]}`. Do not skip a pair or return prose outside the JSON object.

Build sketches used one of two JSON schemas. The expanded schema contained `data`, `architecture`, `loss_and_training` and `evaluation`. The compact schema combined the first three fields in `implementation_summary` and retained `evaluation`. Both schemas also included a short `name` and `tagline`.

Build-sketch construction

You are a computational-biology research engineer converting one graph-grounded Problem–Mechanism candidate into a concrete build sketch. Stay within the supplied research topic and use only choices defensible from the Problem, Mechanism, linked Effects and paper provenance.

The sketch must be grounded, concrete, topic-bounded and explicit about uncertainty. State the input data and required annotations; the architecture or analytical procedure that instantiates the Mechanism; the loss, objective or fitting procedure; and an initial evaluation containing appropriate public datasets, metrics and baseline families. Do not invent a required modality, benchmark or loss that the supplied evidence does not support. When a choice is genuinely underdetermined, select one implementation and name a plausible alternative instead of presenting the choice as settled.

Return strict JSON. If the evidence is insufficient for a concrete sketch, return `INSUFFICIENT_SPEC` and one reason.

MCG candidate ranking

Score every supplied candidate, without pre-filtering or omission, on five integer-valued criteria from 0 to 100:

1. **Gap clarity**: precision and evidential support of the Mechanism-to-Problem mapping.
2. **Impact potential**: importance of the advance if the proposed intervention works.
3. **Feasibility**: implementability and evaluability by a small team using accessible data and baselines.
4. **Methodological fit**: relevance to a recurring methodological need and strength as a methods contribution.
5. **Biological grounding**: whether the design encodes relevant single-cell biology or measurement structure rather than only generic machine-learning machinery.

Use the full scale consistently: 90–100 denotes a fully satisfied criterion; 70–89, strong with at most a minor concern; 50–69, mixed; 30–49, critically weak; and below 30, severely weak or disqualifying. Compute a separate `overall_score` by non-linear synthesis rather than an arithmetic mean. Any criterion below 50 caps the overall score in the corresponding weak band, and any criterion below 30 caps it at 30.

For every input candidate, return its identifier, concise title and description, all five numerical scores, a one- or two-sentence justification for each score, the overall score, the decisive factor and a short holistic rationale. Return strict JSON ordered by overall score. These scores determine internal processing priority; they do not authorize implementation or constitute a scientific-validity judgment.

Blinded temporal relation assessment

Temporal relation judge

ROLE AND BLINDING

You are an independent expert annotator. Judge exactly one blinded query at a time. Do not inspect a private registry, another annotator's responses, retrieval-system identity, paper identity, publication year, rank, retrieval score, Requirement or Articulation. Do not use web search. Base every judgment only on the complete blinded item supplied for that query.

The future reference contains one realized Design and every Effect attached to it. Each historical candidate contains one canonical historical Design and every Effect attached to that occurrence. Judge each candidate independently against the future realized Design-to-Effects reference. Effects provide causal evidence; they are not sufficient grounds for equating different operations.

RELATION LABELS

Assign exactly one relation and its fixed integer grade:

- `SAME_CORE` (grade 3): the same central causal operation acts on the same kind of object and creates the same relevant causal intermediate. Implementation and parameter differences are minor.
- `SPECIALIZES` (grade 2): the future Design retains a reusable central historical operation but adapts, extends or surrounds it with problem-specific machinery. Name both the shared operation and the future-specific change.
- `RELATED_FAMILY` (grade 1): a broad technical family, representation, task or desired Effect is shared, but the central operations differ or an essential operation is absent.
- `DIFFERENT` (grade 0): the pair is mechanistically unrelated, is expressed at the wrong analytical level or modality, or has no plausible causal connection.

Shared vocabulary, task, input modality or outcome alone is insufficient. When evidence is insufficient, choose the lower grade. Do not rank candidates against one another and do not force any candidate to be compatible.

OUTPUT

Return every candidate exactly once. For each candidate return `candidate_key`, `relation`, integer `grade`, `confidence` from 0 to 100, `shared_core_operation` (empty only for `DIFFERENT`), `key_difference`, and one concise, concrete rationale. Return strict JSON only and never shorten or omit supplied text.

Candidate method-hypothesis formulation

The following six boxes report the load-bearing instructions for the proposer sequence. Repetitive examples and runtime serialization wrappers were shortened, while retaining the scope constraints, reasoning order and final record schema.

Proposer: persona

You are a senior bioinformatics researcher specializing in computational tool development, targeting journals such as *Bioinformatics*, *Genome Biology* and *Nucleic Acids Research*, or conferences such as ISMB and RECOMB.
Strong bioinformatics methods papers solve one clear biological or computational problem that practitioners face. The method should be simple enough to explain in one figure, use three to five established public datasets, compare against four to six strong baselines under a common protocol, isolate two or three contributions through ablation, and be reproducible. Avoid stacking many techniques, claiming a foundation or universal framework without corresponding scale, pursuing marginal architecture changes, presenting a purely machine-learning exercise, or requiring an impractical ablation matrix.
Start from a concrete computational problem. Ground every design choice in an established technique or an explicit knowledge gap. Name specific algorithms, architectures, objectives and evaluation resources. Prefer a focused, technically defensible contribution over an unnecessarily complex design.

Proposer step 1: problem framing

Based on the supplied research direction, formulate a focused biological motivation and computational problem. The proposal must remain within the given topic and modality.

SCOPE CONSTRAINT

The proposal must address one clear task, be trainable and evaluable end to end on several public datasets, fit within a single-project implementation, and contain a mechanism that can be explained in one figure and tested with a compact ablation. Do not propose a foundation model, universal framework, multi-task suite, atlas-scale pretraining project or language-model-style pretraining over single-cell corpora.

EXECUTION CONTRACT

The method consumes `.h5ad` files containing matrices and metadata. It returns standard embeddings, labels or per-cell scores in an `.h5ad` object. The environment uses standard Python single-cell and machine-learning libraries. Dataset preparation may create splits, masks, modality drops and deterministic subsamples, but it does not depend on new raw sequencing processing, genome files or external annotation databases. If a proposal falls outside this contract, reformulate it at the matrix level or provide a standard downstream output before continuing.

PROBLEM-VARIANT SEARCH

Prefer an under-served variant of a well-studied problem. Identify (1) the standard task, (2) the assumption that makes it tractable, (3) the realistic setting in which the assumption fails, (4) existing baselines that remain runnable in that setting and (5) the unaddressed variant. This pattern preserves both novelty and evaluability.

MULTIPLE LENSES

Before selecting a direction, record one concrete observation from each lens: the single-cell workflow and its bottleneck; the biological constraint; the information that is present or missing; a matching statistical or machine-learning inductive bias; and a structurally related solution from an adjacent field.

BIOLOGICAL CHALLENGE ANALYSIS

List at least three specific biological or measurement challenges. Explain what makes the data distinctive, what information current workflows discard and which assumptions current methods oversimplify. Use these challenges to drive the subsequent mechanism design.

Write one paragraph for `MOTIVATION` and one for `COMPUTATIONAL PROBLEM`. The latter must state the input, desired output and the exact failure of existing methods. Before returning, verify that the proposal targets one task, remains within a tractable data scale and does not drift toward a universal or foundation-model claim.

Proposer step 2: literature mining and limitations

Using the motivation, computational problem and paper claim, identify the concrete limitations that prevent existing methods from satisfying the claim. Name methods and their specific failure modes rather than listing generic weaknesses.

For each supplied paper, extract (1) the key technique and the task or metric it improves, (2) its author-reported limitation or failure mode and (3) any technique not yet applied to the target problem. From the RPKG evidence, identify unaddressed Problem nodes, Mechanism nodes not yet connected to the target Problem, and relations that document exploitable limitations.

Treat an RPKG path as a compact causal account. A path such as `Problem` → `motivates` → `Mechanism` → `achieves` → `Effect` → `enables` → `Task` provides a candidate skeleton, but it must not be interpreted mechanically. Prefer the two or three most relevant paths by their explicit content. Prioritize paths that contain explicit `limits` or `motivates` relations and are supported by strong provenance. When named motifs are supplied instead, examine shared limitations, unexplored Problem–Mechanism combinations, transferred mechanisms, complete causal chains, unaddressed problems and unused mechanisms.

Synthesize one or two specific gaps that no existing method addresses, state what current approaches oversimplify, and identify inferior or redundant directions to avoid. Then state one testable novelty seed that combines an RPKG gap with a paper-derived finding. Cite paper identifiers and RPKG node names in the working record.

Proposer step 3: method proposal

Propose a method that directly addresses at least two biological challenges from step 1. It must have its own central mechanism rather than assemble existing tools into a pipeline. A new formalism, an enforced structural property, or a problem reformulation is acceptable; a thin wrapper around established encoders, losses or post-processing modules is not.

NOVELTY-CLASS GATE

Classify the candidate before finalizing it:

- **Class A**: a mathematical formalism imported from another field is the central object. Name the formalism, source field and prior problem.
- **Class B**: a new invariance, equivariance or symmetry is architecturally enforced. State it mathematically.
- **Class C**: the problem is reformulated so that named existing methods become explicit special cases, or so a previously disconnected solution class becomes applicable.
- **Class D**: known mechanisms are combined without a new formalism, structural constraint or reformulation. Use this only after considering concrete Class A–C alternatives.

For every non-selected class, name one concrete candidate and a concrete reason for rejection. For a selected Class A–C mechanism, apply the load-bearing test: if removing the claimed formalism, structural constraint or reformulation leaves the core claim intact, it is decorative and the candidate must be revised or reclassified. For a Class D candidate, state the causal chain connecting its parts; independent regularizers do not constitute mechanistic synergy.

FEASIBILITY AND OUTPUT

Use at most two or three core components and two or three objectives. The method should be implementable within one focused project. Return the following sections:

- `CORE INNOVATION`: one key technical insight linked to the most important challenge.
- `TECHNICAL APPROACH`: the causal role of every design choice.
- `KEY COMPONENTS`: two or three components, the challenge addressed by each and the consequence of removing it.
- `CHALLENGE-COMPONENT MAPPING`: an explicit mapping from each addressed challenge to a component.
- `NOVELTY CLAIM`: one evidence-grounded sentence stating what the candidate adds beyond retrieved work.
- `LITERATURE GROUNDING`: a paper finding or RPKG relation supporting each component, with unsupported elements identified as hypotheses rather than evidence.

Proposer step 4: claim-aligned benchmark planning

Begin by identifying one evaluation dimension, task or scenario in which a structural property of the proposed method should provide a distinctive advantage. A generic expectation of better performance on a metric already optimized by strong baselines is insufficient. State why the selected baselines cannot obtain the same property and what result would test it.

Then list five or six public datasets spanning relevant platforms, tissues, species or scales; four to six task-relevant baselines covering the major approach categories; and two to four primary metrics that directly test the claim. Add a runtime and memory comparison, sensitivity analysis for two or three key hyperparameters, one expected result tied to the distinctive mechanism, the most suitable venue and the prior work on which the candidate builds. Use real biological data as the principal evidence; simulations may be supporting analyses. Keep this stage as a concise evaluation outline rather than a detailed experiment protocol.

Proposer: consolidation

Consolidate the previous outputs into one internally consistent method-hypothesis record. The motivation must lead to the computational problem, the limitations must motivate the proposed mechanism, and the evaluation must test the stated claims. Before returning, verify that the core mechanism can be explained in one sentence; no more than three components are present; each component maps to a biological challenge; the problem reflects a practitioner-relevant need; public data and standard metrics can evaluate it; and the planned implementation is tractable. Revise the record when any check fails.
Return only one JSON object:

```
{
  "biological_challenges": ["challenge 1", "challenge 2"],
  "motivation": "...",
  "computational_problem": "...",
  "limitations_of_existing_methods": "...",
  "proposed_method": {
    "core_innovation": "...",
    "technical_approach": "...",
    "key_components": ["...", "..."],
    "challenge_component_mapping": {
      "challenge 1": "component that addresses it",
      "challenge 2": "component that addresses it"
    }
  },
  "evaluation_plan": {
    "benchmarks": ["dataset_name_1", "dataset_name_2"],
    "baselines": ["method_1", "method_2"],
    "metrics": ["metric_1", "metric_2"]
  },
  "expected_improvement": "one sentence",
  "target_venue": "...",
  "related_work": ["..."]
}
```

Evaluation-plan fields contain names rather than detailed execution instructions.

Blinded candidate-formulation evaluation

Secondary five-dimension model rating

You are an independent evaluator of a candidate computational method hypothesis for single-cell analysis. Evaluate only the anonymized candidate text supplied in the current item. Do not infer its generation condition, reward polished wording by itself, compare it with remembered candidates, browse for additional evidence, or treat the task as peer review. The scores characterize the written proposal under a fixed rubric; they do not establish scientific validity, empirical feasibility or whether a researcher should proceed.

Assign one integer score from 1 to 10 for each of the following dimensions, using the anchors in Supplementary Table S5:

- **Novelty**: distinctness from nearby existing approaches for the stated Problem.
- **Significance**: how much the proposal would advance the stated Problem if its claims held.
- **Mechanistic specificity**: concreteness of the causal operation, components, design choices and actionable steps.
- **Biological grounding**: use of relevant scRNA-seq biology, measurement properties and data structure.
- **Feasibility**: implementability and benchmarkability with identifiable tools, data and metrics.

Judge each dimension separately. Do not raise a score because another dimension is strong. When information needed for a dimension is absent, score the candidate according to what is written rather than filling the gap with an assumed implementation. Return strict JSON only:

```
{
  "novelty": {"score": 1, "rationale": "..."},
  "significance": {"score": 1, "rationale": "..."},
  "mechanistic_specificity": {"score": 1, "rationale": "..."},
  "biological_grounding": {"score": 1, "rationale": "..."},
  "feasibility": {"score": 1, "rationale": "..."}
}
```

Human forced-choice instruction

For the same computational Problem, read the three anonymized candidate method hypotheses shown in randomized order. Select exactly one candidate that you would be most willing to pursue for further method development. Base the choice on the candidate as a whole, including whether its Mechanism addresses the stated failure, whether the design is sufficiently concrete to investigate and whether its proposed evaluation could inform the next research decision. Do not attempt to identify the generating condition. A tie, abstention or “none of these” response is not available. This is a relative choice within the triplet, not a declaration that the selected candidate is scientifically valid or ready for implementation.

Supplementary Note 6: Review-style assessment of the research records

The four research records for AmbientCap, scOSCAR, CellWDA and CONFIDE-VI were assessed by the same six-researcher panel that completed the blinded candidate-selection study. The panel comprised three doctoral-level and three master's-level researchers acting as simulated reviewers under criteria adapted from the reviewer guidelines of the journal *Bioinformatics*. Reviewers were informed that the records had been assembled by scHypothesizer. Their comments were collected after the records and empirical results had been finalized and were not used to revise any reported hypothesis, implementation or benchmark result.

For every record, the reviewers identified substantive, case-specific strengths and weaknesses that could be related to the proposed Mechanism and its reported evidence (Supplementary Table S7). Despite differing assessments of the individual candidates, the panel agreed that the records exposed enough of the claim, evidence and empirical boundary to provide informative starting points for subsequent researcher judgment. This agreement concerned the usefulness of the records for deciding on further development.

Table S7 | Representative strengths and weaknesses identified in the four research records. Entries summarize recurring observations across the six simulated reviewers and concern the information available for subsequent judgment, rather than an overall ranking of the candidates.

Record	Reviewer-identified strengths	Reviewer-identified weaknesses
AmbientCap	An interpretable gene–cluster decision rule, a clear two-component design and evidence connecting rare-cell preservation to the proposed operation.	The strength of the rare-cell evidence depended on the benchmark construction, while comparisons with several established ambient-RNA methods remained limited.
scOSCAR	A well-motivated open-set mapping problem, calibrated set-valued outputs and interpretable class-level evidence for unsupported cells.	Evidence was stronger for broad novel populations than for near-lineage novelty, and the breadth of shift settings was limited.
CellWDA	A clear partial-overlap integration problem, an interpretable per-cell support gate and benchmark evidence tied to selective alignment.	Several results were close to comparator performance, and the available assessment provided limited evidence about robustness across repeated runs and settings.
CONFIDE-VI	Transparent reporting of negative findings, a traceable per-claim assessment and measurable behaviour of the density-stratified scaffold.	The central alignability operation performed near chance, and several dependent claims were not supported by the reported results.

Supplementary Data

The files accompanying the candidate-hypothesis, human-selection and secondary-rating analyses are listed below.

Supplementary Data 1: consolidated method hypotheses

`SuppData1_consolidated_hypotheses.json`. Thirty complete scHypothesizer candidate records, comprising ten records for each of Ambient decontamination, rare-cell-preserving Integration and uncertainty-aware Reference mapping. The human-readable versions appear in Supplementary Note 4.

Supplementary Data 2: secondary blinded model ratings

- `SuppData2_blinded_model_scores.xlsx`
- `SuppData2_blinded_model_scores_by_direction.csv`

These files provide the direction-level summaries underlying Supplementary Fig. S1 and Supplementary Tables S4 and S6. The accompanying `README.md` maps the condition names used in the files to those used in the manuscript.

Supplementary Data 3: anonymized human forced-choice decisions

`SuppData3_human_forced_choice_decisions.csv`. An anonymized long-form table containing the 180 forced-choice decisions underlying Supplementary Table S1. Each row records a reviewer alias, candidate-set identifier, topic, blinded option selected and the condition revealed after all choices were complete. Original reviewer identifiers and response timestamps are not included.

References

[1] Matthew D Young and Sam Behjati. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *GigaScience*, 9(12):giaa151, November 2020. ISSN 2047-217X. doi: 10.1093/gigascience/giaa151. URL https://doi.org/10.1093/gigascience/giaa151.

[2] Shiyi Yang, Sean E. Corbett, Yusuke Koga, et al. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biology*, 21(1):57, March 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-1950-6. URL https://doi.org/10.1186/s13059-020-1950-6.

[3] Weijian Wang, Yihui Cen, Zezhen Lu, et al. scCDC: a computational method for gene-specific contamination detection and correction in single-cell and single-nucleus RNA-seq data. *Genome Biology*, 25(1):136, May 2024. ISSN 1474-760X. doi: 10.1186/s13059-024-03284-w. URL https://doi.org/10.1186/s13059-024-03284-w.

[4] Marijn Berg, Ilya Petoukhov, Inge van den Ende, et al. FastCAR: fast correction for ambient RNA to facilitate differential gene expression analysis in single-cell RNA-sequencing datasets. *BMC Genomics*, 24(1):722, November 2023. ISSN 1471-2164. doi: 10.1186/s12864-023-09822-3. URL https://doi.org/10.1186/s12864-023-09822-3.

[5] Chenling Xu, Romain Lopez, Edouard Mehlman, et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1):MSB20209620, January 2021. ISSN 1744-4292. doi: 10.15252/msb.20209620. URL https://doi.org/10.15252/msb.20209620.

[6] Vladimir Yu Kiselev, Andrew Yiu, and Martin Hemberg. scmap: projection of single-cell RNA-seq data across data sets. *Nature Methods*, 15(5):359–362, May 2018. ISSN 1548-7105. doi: 10.1038/nmeth.4644. URL https://doi.org/10.1038/nmeth.4644.

[7] Yi-Xuan Xiong and Xiao-Fei Zhang. scDOT: enhancing single-cell RNA-Seq data annotation and uncovering novel cell types through multi-reference integration. *Briefings in Bioinformatics*, 25(2):bbae072, March 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae072. URL https://doi.org/10.1093/bib/bbae072.

[8] Yi-Xuan Xiong, Meng-Guo Wang, Luonan Chen, et al. Cell-type annotation with accurate unseen cell-type identification using multiple references. *PLOS Computational Biology*, 19(6):1–24, June 2023. doi: 10.1371/journal.pcbi.1011261. URL https://doi.org/10.1371/journal.pcbi.1011261.

[9] Tim Stuart, Andrew Butler, Paul Hoffman, et al. Comprehensive Integration of Single-Cell Data. *Cell*, 177(7):1888–1902.e21, June 2019. ISSN 0092-8674. doi: 10.1016/j.cell.2019.05.031. URL https://doi.org/10.1016/j.cell.2019.05.031.

[10] Marcos López-De-Castro, Alberto García-Galindo, José González-Gomariz, et al. Conformal inference for reliable single cell RNA-seq annotation. *Bioinformatics*, 41(10):btaf521, October 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf521. URL https://doi.org/10.1093/bioinformatics/btaf521.

[11] Romain Lopez, Jeffrey Regier, Michael B. Cole, et al. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, December 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0229-2. URL https://doi.org/10.1038/s41592-018-0229-2.

[12] Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, 37(6):685–691, June 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0113-3. URL https://doi.org/10.1038/s41587-019-0113-3.

[13] Ilya Korsunsky, Nghia Millard, Jean Fan, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, December 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0619-0. URL https://doi.org/10.1038/s41592-019-0619-0.

[14] Xiaokang Yu, Xinyi Xu, Jingxiao Zhang, et al. Batch alignment of single-cell transcriptomics data using deep metric learning. *Nature Communications*, 14(1):960, February 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-36635-5. URL https://doi.org/10.1038/s41467-023-36635-5.

[15] Krzysztof Polański, Matthew D Young, Zhichao Miao, et al. BBKNN: fast batch alignment of single cell transcriptomes. *Bioinformatics*, 36(3):964–965, February 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz625. URL https://doi.org/10.1093/bioinformatics/btz625.

[16] Massimo Andreatta and Santiago J Carmona. STACAS: Sub-Type Anchor Correction for Alignment in Seurat to integrate single-cell RNA-seq data. *Bioinformatics*, 37(6):882–884, March 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa755. URL https://doi.org/10.1093/bioinformatics/btaa755.