ARTIFACT INJECTION TO BOOST ANOMALY DETECTION IN KIDNEY SINGLE-CELL OMICS

RESEARCH PLAN • DRAFT FOR DISCUSSION

Hayat Lab, Uniklinik RWTH Aachen
Draft for internal discussion; author list to be confirmed

31 October 2026

ABSTRACT

In kidney single-cell and single-nucleus RNA-seq, a cell that looks abnormal may be diseased, or its abnormality may come from how it was measured: doublets, ambient RNA, low-quality cells or dissociation stress. The two look alike, current methods flag both as anomalous, and we found no dataset with known answers that tests whether they can be told apart. We propose a pipeline that turns any single-cell dataset into a labelled anomaly-detection dataset by injecting known technical artifacts, following how each artifact arises, and recording the type, strength and source cells of every injected cell. The labelled data are used for three purposes: (i) to fine-tune ("boost") anomaly detectors and open single-cell foundation models so that they separate artifacts from disease; (ii) to audit whether a model mistakes artifacts for disease; and (iii) to audit how many artifacts a dataset contains. Injected artifacts are used for training only. External validation uses only cells from public data whose truth is known: genotype-identified doublets from pooled donors, the same donor measured with several assays, expert-labelled diseased cells, and healthy donors from a second laboratory. This document describes the design; no results are reported yet.

Keywords anomaly detection · single-cell RNA-seq · kidney · doublets · ambient RNA · synthetic anomalies · foundation models

1 Introduction

The question. In kidney single-cell and single-nucleus data, some cells do not match healthy tissue. They may be diseased, or their abnormality may come from how they were measured, and the two look alike. The two overlap: one of the criteria Lake et al. [1] used for possible altered states is "low genes detected, a high number of mitochondrial transcripts", which are also the signs of a low-quality cell. Our reading (as of October 2026) is that current methods do not set out to separate the two, and that no dataset with known answers exists to test whether they can tell the two apart, or how many false alarms they raise. We therefore cannot say how far findings built on these anomalous cells can be trusted. This plan goes one step further and asks: *can artifacts that we create teach models to tell the two apart?*

Why it matters. Measurement effects have produced false biology: in human brain single-nucleus data, nuclei previously annotated as "immature oligodendrocytes" turned out to be glia contaminated with neuronal ambient RNA [2]. Background contamination in kidney single-nucleus data can be high: in mixed-strain mouse kidney, single-nucleus samples from frozen tissue had clearly more background than single-cell samples from the same material (35% vs 11%, 17% vs 3%) [3], and frozen kidney biopsies can only be profiled by single-nucleus sequencing [4]. Disease and measurement signals also overlap. The kidney atlas of Lake et al. [1] flagged candidate altered (injury) states partly by few genes detected and many mitochondrial transcripts, which in our reading are also the usual signs of a low-quality cell; altered states also occur in healthy reference tissue, which the authors attribute possibly to ischaemia during acquisition or normal ageing. From the public label table we counted that 26% of nuclei from healthy donors are altered (22,774 of 88,460).

Why it is unsolved. As of October 2026, four gaps remain in our reading. Artifact benchmarks have no disease axis: doublets and ambient RNA have their own benchmarks, but each one we found covers one artifact and no disease states, for example the ambient-removal benchmark of Cargnelli et al. [5]. Disease-state detection does not separate artifacts: methods that find altered states against a healthy reference [6] aim at disease states and do not judge whether a flagged cell is an artifact; doublets and low-quality cells are only removed in preprocessing, and the technical source of false positives the authors warn about is residual batch effects. There is no error control: annotation pipelines with a reject option "do not provide statistical guarantees" [7], and conformal methods can bound the false-alarm rate [8] only when reference and query are exchangeable; in the single-cell conformal study the guarantee failed for queries from lung-tumour patients, and we found no systematic test on kidney data. Finally, there is no test bed: the general anomaly-detection benchmarks we examined contain no single-cell expression data, and we found no anomaly-detection benchmark for kidney omics.

Has this direction been done? The search cannot be exhaustive, so we state findings "to our knowledge". The parts exist, each for one artifact: simulating doublets and training a detector on them is standard [9–11]; an R package simulates damaged cells [12], with no accompanying paper; and ambisim simulates ambient molecules at read level to benchmark demultiplexing [13]. The problem has been named: mapQC warns that a poor mapping can result in "falsely reporting technical artifacts as disease" [14], but it addresses batch and mapping quality, not injection. A zero-shot evaluation found scGPT and Geneformer embeddings often behind simpler methods [15], but it tested batch effects; we found no study that tests foundation models against doublets, ambient RNA or low-quality cells. We found no tool that injects several artifact types into any real dataset with recorded answers and uses them to train models to tell artifact from disease.

Evidence and warnings from image anomaly detection. DRAEM, trained on synthetic defects, raised image-level AUROC on MVTec AD from the previous best of 95.5 to 98.0, and its authors note that the synthetic defects need not look realistic [16]. However, across 420 tasks, Yoo et al. [17] found that when the augmentation does not match how real anomalies arise, self-supervised training can even hurt. In single cells, detectors trained on simulated doublets reach a best mean AUPRC of only 0.537 on 16 datasets with real labels [18]. Two hard rules follow: *inject artifacts by their cause, and validate externally on real data only.* The three artifacts can be simulated to very different degrees (Table 3, Appendix A): doublets have a mature recipe, ambient RNA can reuse existing removal models in reverse, and low-quality cells can only be approximated.

Planned contributions. We plan an open-source pipeline that injects artifacts into any single-cell dataset (AnnData) and records the type, strength and source cells of every injected cell; a boosting recipe that trains models on the injected artifacts so that they separate healthy, diseased and artifact cells; an external validation set with fixed tasks, metrics and donor-level splits; a results table for each omics model with its anomaly score, run with and without the pipeline, reported per artifact type and with a leave-one-artifact-out protocol; a dataset audit that applies the trained detector to public atlases to estimate their artifact share and flag suspicious altered clusters for expert review; and, as a second stage, an extension to RNA and ATAC. Raw data are not redistributed; code, split files, injection records and derived labels are released (for the second cohort, only once its data licence is confirmed).

2 Research questions

The plan asks one main question and three follow-up questions; each is answered by the tasks of the external validation (Section 3.4). **RQ1, boosting:** does training on injected artifacts help models detect *real* artifacts and diseased cells, compared with the same models trained on healthy cells only? This is answered by tasks T1 to T3, always against a floor that uses only gene count and mitochondrial share. **RQ2, model audit:** do current models, and single-cell foundation models in particular, place artifacts close to diseased cells, that is, would they mistake an artifact for disease? **RQ3, robustness:** do the false-alarm thresholds still hold when the assay changes or the data come from another laboratory? This is answered by tasks T2b and T4. **RQ4, dataset audit:** applied to public kidney atlases, how large a share of cells does the trained detector flag as artifacts, and which altered clusters become suspicious and are confirmed or rejected by clinical experts?

3 Method

3.1 Overview

Figure 1 and Table 1 summarise the method. Artifact injection creates known artifacts in real data and records every one. Boosting trains an artifact discriminator on these labelled data, on top of omics models and anomaly scores that

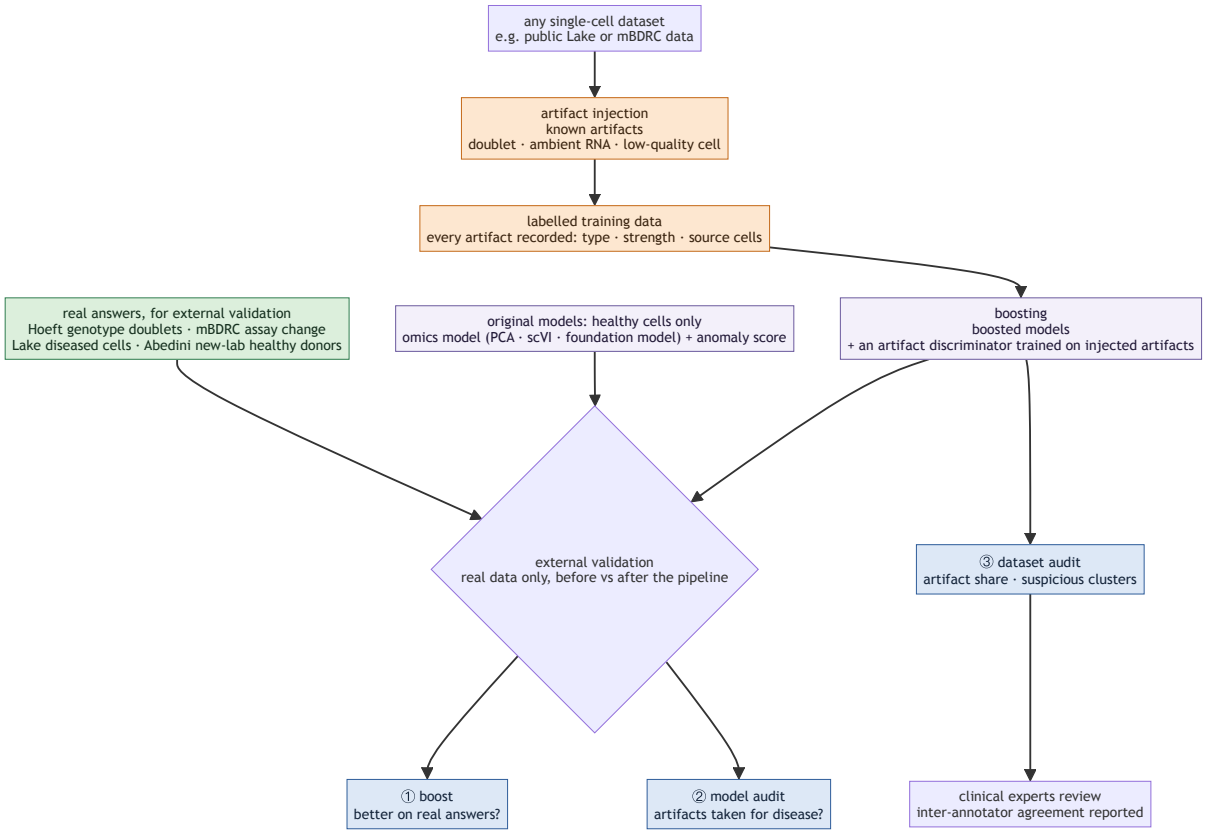


Figure 1: Design overview. Orange: injected artifacts and the labelled data they produce. Purple: models, run twice (original, and boosted with an artifact discriminator). Green: real answers used only for external validation. Blue: the three uses.

are also kept in their original form. External validation then compares the original and the boosted models only on real cells whose truth is known. The comparison answers use (i) and (ii); the boosted detector applied to a dataset gives use (iii), whose suspicious clusters go to clinical experts.

Table 1: Overview of the method.

Stage	What it does	Data	Output
Artifact injection	create doublets, ambient RNA and low-quality cells in real data, recording every answer	healthy cells [1, 19]	labelled training data
Boosting	train models on these answers to separate healthy, diseased and artifact cells	injected data	boosted models
External validation	check only on real cells whose truth is known, before vs after boosting	Hoeft et al. [4], Acera-Mateos et al. [19], Lake et al. [1], Abedini et al. [20]	whether the pipeline works

3.2 Artifact injection

Which artifacts. The main line covers three artifacts, starting with doublets: they have the strongest precedent and a ready real answer, so they are the best way to run the whole loop (inject, boost, validate externally, report) end to end; adding ambient RNA and low-quality cells then only swaps the injection module. All three are needed because each one can imitate a different "disease" (Table 2; the second column is our inference).

How each artifact is injected. The three artifacts can be simulated to very different degrees (Table 3); full recipes, parameters and sources are in Appendix A.

Table 2: Artifacts, what they can be taken for, and the plan.

Artifact	Could be taken for	Why	Plan
Low-quality cell	the degenerative state: few genes, high mitochondrial share	the criteria of Lake et al. [1] for possible altered states include "low genes detected, a high number of mitochondrial transcripts"	main line, third
Doublet	an intermediate state expressing two cell types' genes	doublets "suggest the existence of intermediate cell states" [9]	main line, first
Ambient RNA	marker genes in the wrong cell, e.g. proximal-tubule (PT) genes elsewhere	the PT-only gene Slc34a1 is found in a high fraction of other cells in mouse kidney [3]	main line, second
Dissociation stress	injury (FOS, JUN up)	mouse evidence [21]	not injected: no precedent; main data are single-nucleus
Batch or assay change	—	not a per-cell artifact	tested on real data (T2b, T4)

Table 3: Injection recipes at a glance (details in Appendix A).

Artifact	Recipe	Maturity	Main limit
Doublet	sum the raw counts of two cells of different types, shrinking a quarter of the library sizes, as in scDblFinder [11]; precedents [9, 10]	mature	simulated doublets are easier than real ones [18]; same-type doublets are invisible
Ambient RNA	mix a background profile into each cell at a Beta-distributed fraction, reusing removal models in reverse [22–24]; fractions calibrated on Janssen et al. [3]	needs a background profile	the background must be approximated (no empty droplets in the main human data), and ambient RNA is continuous, so the artifact threshold is our choice
Low-quality cell	single cells: remove cytoplasmic RNA [12, 25]; single nuclei: add cytoplasmic material [26, 27]	approximate only	no validated simulator and no real answer; high mitochondrial share is normal for PT [28]

Design rules. The recipes follow six rules that we draw from the evidence above. Artifacts are injected by their cause rather than for realism: two cells in one droplet, background RNA mixed in, cytoplasmic RNA lost in single cells or cytoplasmic material added in single nuclei. Training and validation are kept apart: injected artifacts are used for training, and external validation uses real data or at least a different recipe. One artifact type is left out of training at a time, for example training on doublets and ambient RNA and testing on low-quality cells, to see whether a model has learned what artifacts share or only memorised a recipe. Results are reported per artifact type, since rankings are expected to change between types: on tabular data, ADBench found no unsupervised method statistically better than the others overall, with performance depending on whether a method's assumptions match the anomaly type [29]. The amount and the recipe are varied, both the ratio of synthetic to normal cells and one recipe versus a mix. Finally, recipes are set per assay and cell type: low-quality recipes differ for single cells and single nuclei, the mitochondrial share is judged per cell type, and proximal tubule is scored separately.

Second stage: ATAC. RNA comes first; ATAC is added once the RNA stage works (suggested, not decided), using existing ATAC precedents (Appendix F).

3.3 Anomaly detection and boosting

Boosting uses two kinds of models, chained together: an *omics model* turns each cell into a vector, and an *anomaly-detection model* scores the vectors (Table 4). The main line uses RNA models, since the main data are single-nucleus RNA; ATAC is stage 2.

Before and after the pipeline. The pipeline adds one thing: on the omics model's vectors, an *artifact discriminator*, a simple classifier trained on the injected artifacts, answers whether a cell looks like an artifact (Table 5).

Training has only healthy and artifact labels, so disease is found by exclusion, and diseased cells appear only in external validation. For a fair comparison, the runs with and without the pipeline use the same vectors and the same anomaly

Table 4: The two kinds of models.

	Omics model	Anomaly-detection model
What it does	describes a cell as a vector	scores how far a cell is from normal
Input → output	gene counts → a vector	vectors → one anomaly score per cell
Candidates	HVG + PCA; scVI [30]; foundation models Geneformer V2 [31] and scGPT [32]	kNN distance, Isolation Forest, autoencoder reconstruction error

Table 5: What the pipeline changes.

	Without the pipeline	With the pipeline
Training data	healthy cells only	healthy cells + injected artifacts with labels
What it learns	what normal looks like	what normal and artifacts look like
Output per cell	an anomaly score	anomaly score + artifact-like or not
Call	high score = anomalous	low → normal; high and artifact-like → artifact; high, not artifact-like → possibly disease
Disease vs artifact?	no	yes

score and differ only in the artifact discriminator: if every omics model improves, the pipeline works; if only some do, the gain depends on the omics model; if none do, the injection must be revised. The pipeline acts first on the anomaly-detection step, which is cheap, and later, at far higher cost, by fine-tuning the omics model itself on the artifacts; both follow image anomaly detection, which trains with known outliers [33] or trains a discriminator on synthetic defects [16]. The first round (suggested, not decided) uses kNN scores on frozen vectors, and its simplest combination, PCA with kNN, conformal thresholds and the artifact discriminator, serves as our reference method.

Floor and reference points. A QC floor scores cells by gene count and mitochondrial share only, without training; since the disease labels of Lake et al. [1] partly rest on the same two signs, a method that cannot beat it has learned only "poor quality". Specialist doublet detectors [11, 34] are already trained on simulated doublets, so they serve as reference points on T2 rather than in the before-and-after comparison. SoupX and DecontX estimate contamination per sample and per cell and then correct counts [22, 23]; they do not decide whether a cell is an artifact or diseased, so we borrow their models for ambient injection, with DecontX's per-cell estimate as a reference point on the ambient task. Finally, the model audit (use ii) is a separate experiment without boosting: do injected artifacts sit close to real diseased cells in an omics model's vector space?

More details. Foundation models: no pretraining from scratch, which would need tens of millions of cells [35]; Geneformer V2 [31] and scGPT [32] first, as frozen vectors and later fine-tuned; they must be compared with simple methods, which they often trailed zero-shot [15] (candidates in Appendix B). Uncertain cells: donor-grouped cross-validation with out-of-fold scores [36], calibration [37], conformal thresholds [38] that lose their guarantee in a new laboratory [7], and expert review without using uncertain cells as labels [39], since a low score alone cannot say whether a cell is an artifact or early disease [40] (Appendix C). Whether to build a new model later is discussed in Appendix E.

3.4 External validation

Models are trained on artifacts we create; testing them on created artifacts would only show that they memorised the recipe. Validation therefore uses only cells from real data whose truth is known, never touched during training or tuning: an external validation. With no wet laboratory, the truth comes from public data whose experimental design already carries the answer, or from expert judgement (Table 6). Every check reports two numbers, without and with the pipeline, plus leave-one-artifact-out results (Table 7).

A known trap. Using the altered labels of Lake et al. [1] directly as disease ground truth is partly circular, in our reading: the authors also used injury markers and AKI/CKD enrichment, not only QC features, but in the single-nucleus data the largest altered group (degenerative) has less than half the genes of normal nuclei and about five times the mitochondrial fraction, so a scorer that looks only at QC metrics could score well. A QC-only baseline and per-state reporting are therefore required. A teaching pipeline in our onboarding material already shows a first sign of this: on degenerative nuclei a healthy-reference kNN score reaches AUROC 0.76–0.81 and a gene-count baseline 0.76–0.85 over five donor splits, with kNN worse in four of five (a teaching result, not a result of this project).

Table 6: Where the truth comes from.

Data	Known for sure	Why	Reliability
Hoeft et al. [4]	which droplets are real doublets	four people pooled in one run; a droplet with two people's DNA must hold two cells	fixed by design; a lower bound
Acera-Mateos et al. [19]	differences that are not disease	the same healthy person measured with several assays	fixed by design
Lake et al. [1]	which cells are diseased	the authors labelled altered states	expert judgement; weakest
Abedini et al. [20]	cells that should not be flagged	healthy donors from another laboratory	our design choice

Table 7: Tasks. α is the target false-alarm rate.

Task	Question	Truth	Metric
T1	find diseased cells	altered states [1] (expert judgement)	AUROC, AUPRC; against the QC floor
T2	find real measurement artifacts	genotype doublets [4] (fixed by design, a lower bound); held-out injected types as a secondary check	AUROC, AUPRC
T2b	is an assay change flagged as anomalous?	same donor across assays [19] (fixed by design)	false-alarm rate $\leq \alpha$
T3	separate disease from artifact	T1 and T2 together	macro-F1, confusion matrix
T4	do false alarms stay controlled in a new lab?	every call on a healthy donor of a second cohort [20] counts as a false alarm (our design choice; healthy tissue also carries altered cells)	observed rate vs α

Making the disease answer more reliable. The disease labels are expert judgement, the weakest link. Clinical experts (nephrologists, renal pathologists) can strengthen them in two ways: (A) re-annotate each altered cluster as likely injury, likely low quality or unsure, with several independent annotators and their agreement reported; or (B) outline injured regions on the H&E images paired with the public Visium data and check anomaly scores spot by spot. Details are in Appendix D.

4 Datasets

Only public kidney data are used: human first, mouse only for calibration. Table 8 lists what each dataset is used for. Splits are by donor, not by cell, so no person appears in both training and external validation; healthy donors are split into reference, calibration and test sets, and the split files ship with the dataset.

References

[1] Blue B. Lake et al. An atlas of healthy and injured cell states and niches in the human kidney. *Nature*, 619(7970): 585–594, July 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05769-3. URL `http://dx.doi.org/10.1038/s41586-023-05769-3`.

[2] Emre Caglayan, Yuxiang Liu, and Genevieve Konopka. Neuronal ambient RNA contamination causes misinterpreted and masked cell types in brain single-nuclei datasets. *Neuron*, 110(24):4043–4056.e5, December 2022. ISSN 0896-6273. doi: 10.1016/j.neuron.2022.09.010. URL `http://dx.doi.org/10.1016/j.neuron.2022.09.010`.

[3] Philipp Janssen et al. The effect of background noise and its removal on the analysis of single-cell expression data. *Genome Biology*, 24(1), 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-02978-x. URL `http://dx.doi.org/10.1186/s13059-023-02978-x`.

[4] Konrad Hoeft et al. Label-free single-cell RNA multiplexing leveraging genetic variability. *Nature Communications*, 15(1), December 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54270-6. URL `http://dx.doi.org/10.1038/s41467-024-54270-6`.

Table 8: Public datasets and their use.

Dataset	Used in	Role	Caveat
Lake et al. [1]	injection, validation	healthy reference, disease-state labels (T1, T3), base for injection	raw sequencing is controlled, so no empty droplets; state labels and mitochondrial fraction only in the GEO label table
Acera-Mateos et al. [19]	injection, validation	11 of 18 donors measured by two or three assays (T2b); also a base for injection	assay is confounded with tissue core and preservation medium; healthy cortex only; no empty droplets
Abedini et al. [20]	validation	second lab and cohort; healthy controls for T4	ambient RNA and doublets were removed before release, so it cannot test the measurement side; licence not stated
Janssen et al. [3]	injection (mouse)	calibrates injected ambient RNA; files include empty droplets	mouse only; not a test set
Hoeft et al. [4]	validation	real human-kidney doublets and ambient RNA by genotype (T2)	four patients; same-donor doublets are invisible, so the genotype doublets are a lower bound; the public matrix is filtered and must be re-run from FASTQ
Lake et al. [26]	validation (open)	the same person's tissue stored and processed in different ways	under discussion; snDrop-seq, small

[5] Cecilie Bøgh Cargnelli, Jakob Vennike Nielsen, and Jesper Grud Skat Madsen. Benchmarking computational decontamination of ambient RNA. *bioRxiv*, January 2026. doi: 10.64898/2026.01.13.699237. URL http://dx.doi.org/10.64898/2026.01.13.699237.

[6] Emma Dann, Ana-Maria Cujba, Amanda J. Oliver, Kerstin B. Meyer, Sarah A. Teichmann, and John C. Marioni. Precise identification of cell states altered in disease using healthy single-cell references. *Nature Genetics*, 55(11): 1998–2008, October 2023. ISSN 1546-1718. doi: 10.1038/s41588-023-01523-7. URL http://dx.doi.org/10.1038/s41588-023-01523-7.

[7] Marcos López-De-Castro, Alberto García-Galindo, José González-Gomariz, and Rubén Armañanzas. Conformal inference for reliable single cell RNA-seq annotation. *Bioinformatics*, 41(10), September 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf521. URL http://dx.doi.org/10.1093/bioinformatics/btaf521.

[8] Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. Testing for Outliers with Conformal p-values. *The Annals of Statistics*, 51(1):149–178, 2023. doi: 10.1214/22-AOS2244. URL https://doi.org/10.1214/22-AOS2244.

[9] Samuel L. Wolock, Romain Lopez, and Allon M. Klein. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Systems*, 8(4):281–291.e9, April 2019. ISSN 2405-4712. doi: 10.1016/j.cels.2018.11.005. URL http://dx.doi.org/10.1016/j.cels.2018.11.005.

[10] Christopher S. McGinnis, Lyndsay M. Murrow, and Zev J. Gartner. DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. *Cell Systems*, 8(4):329–337.e4, April 2019. ISSN 2405-4712. doi: 10.1016/j.cels.2019.03.003. URL http://dx.doi.org/10.1016/j.cels.2019.03.003.

[11] Pierre-Luc Germain, Aaron Lun, Carlos Garcia Meixide, Will Macnair, and Mark D. Robinson. Doublet identification in single-cell sequencing data using scDblFinder. *F1000Research*, 10:979, May 2022. ISSN 2046-1402. doi: 10.12688/f1000research.73600.2. URL http://dx.doi.org/10.12688/f1000research.73600.2.

[12] Alicen Henning. DamageDetective: Detecting Damaged Cells in Single-Cell RNA Sequencing Data, April 2025. URL http://dx.doi.org/10.32614/CRAN.package.DamageDetective.

[13] Terence Li, Marcus Alvarez, Cuining Liu, Kevin Abuhanna, Yu Sun, Jason Ernst, Kathrin Plath, Brunilda Balliu, Chongyuan Luo, and Noah Zaitlen. The impact of ambient contamination on demultiplexing methods for single-nucleus multiome experiments. February 2025. doi: 10.1101/2025.02.06.636969. URL http://dx.doi.org/10.1101/2025.02.06.636969.

[14] Lisa Sikkema, Mariam Lkandushi, Daniele Scarcella, Amir Ali Moinfar, Jan Engelmann, and Fabian J. Theis. Automated evaluation of single-cell reference atlas mappings enables the identification of disease-associated cell states. *bioRxiv*, May 2025. doi: 10.1101/2025.05.23.655749. URL http://dx.doi.org/10.1101/2025.05.23.655749.

[15] Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26(1), April 2025. ISSN 1474-760X. doi: 10.1186/s13059-025-03574-x. URL http://dx.doi.org/10.1186/s13059-025-03574-x.

[16] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM – A discriminatively trained reconstruction embedding for surface anomaly detection, 2021. URL https://arxiv.org/abs/2108.07610.

[17] Jaemin Yoo, Tiancheng Zhao, and Leman Akoglu. Data Augmentation is a Hyperparameter: Cherry-picked Self-Supervision for Unsupervised Anomaly Detection is Creating the Illusion of Success, 2022. URL https://arxiv.org/abs/2208.07734.

[18] Nan Miles Xi and Jingyi Jessica Li. Benchmarking Computational Doublet-Detection Methods for Single-Cell RNA Sequencing Data. *Cell Systems*, 12(2):176–194.e6, February 2021. ISSN 2405-4712. doi: 10.1016/j.cels.2020.11.008. URL http://dx.doi.org/10.1016/j.cels.2020.11.008.

[19] Mario Acera-Mateos et al. Systematic evaluation of single-cell multimodal data integration enhances cell type resolution and discovery of clinically relevant states in complex tissues. *Genome Biology*, 27(1), March 2026. ISSN 1474-760X. doi: 10.1186/s13059-026-04002-4. URL http://dx.doi.org/10.1186/s13059-026-04002-4.

[20] Amin Abedini et al. Single-cell multi-omic and spatial profiling of human kidneys implicates the fibrotic microenvironment in kidney disease progression. *Nature Genetics*, 56(8):1712–1724, 2024. ISSN 1546-1718. doi: 10.1038/s41588-024-01802-x. URL http://dx.doi.org/10.1038/s41588-024-01802-x.

[21] Elena Denisenko et al. Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. *Genome Biology*, 21(1), 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02048-6. URL http://dx.doi.org/10.1186/s13059-020-02048-6.

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

[23] Shiyi Yang, Sean E. Corbett, Yusuke Koga, Zhe Wang, W Evan Johnson, Masanao Yajima, and Joshua D. Campbell. Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biology*, 21(1), March 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-1950-6. URL http://dx.doi.org/10.1186/s13059-020-1950-6.

[24] Stephen J. Fleming, Mark D. Chaffin, Alessandro Arduini, Amer-Denis Akkad, Eric Banks, John C. Marioni, Anthony A. Philippakis, Patrick T. Ellinor, and Mehrtash Babadi. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nature Methods*, 20(9):1323–1335, August 2023. ISSN 1548-7105. doi: 10.1038/s41592-023-01943-7. URL http://dx.doi.org/10.1038/s41592-023-01943-7.

[25] Tomislav Ilicic, Jong Kyoung Kim, Aleksandra A. Kolodziejczyk, Frederik Otzen Bagger, Davis James McCarthy, John C. Marioni, and Sarah A. Teichmann. Classification of low quality cells from single-cell RNA-seq data. *Genome Biology*, 17(1), February 2016. ISSN 1474-760X. doi: 10.1186/s13059-016-0888-1. URL http://dx.doi.org/10.1186/s13059-016-0888-1.

[26] Blue B. Lake et al. A single-nucleus RNA-sequencing pipeline to decipher the molecular anatomy and pathophysiology of human kidneys. *Nature Communications*, 10(1), 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-10861-2. URL http://dx.doi.org/10.1038/s41467-019-10861-2.

[27] Eloi Schmauch et al. QClus: a droplet filtering algorithm for enhanced snRNA-seq data quality in challenging samples. *Nucleic Acids Research*, 53(1), December 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1145. URL http://dx.doi.org/10.1093/nar/gkae1145.

[28] Ayshwarya Subramanian, Mikhail Alperovich, Yiming Yang, and Bo Li. Biology-inspired data-driven quality control for scientific discovery in single-cell transcriptomics. *Genome Biology*, 23(1), December 2022. ISSN 1474-760X. doi: 10.1186/s13059-022-02820-w. URL http://dx.doi.org/10.1186/s13059-022-02820-w.

[29] Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. ADBench: Anomaly Detection Benchmark, 2022. URL https://arxiv.org/abs/2206.09426.

[30] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15(12):1053–1058, November 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0229-2. URL http://dx.doi.org/10.1038/s41592-018-0229-2.

[31] Christina V. Theodoris et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, May 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06139-9. URL http://dx.doi.org/10.1038/s41586-023-06139-9.

[32] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, February 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02201-0. URL http://dx.doi.org/10.1038/s41592-024-02201-0.

[33] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep Anomaly Detection with Outlier Exposure, 2018. URL https://arxiv.org/abs/1812.04606.

[34] Abha S Bais and Dennis Kostka. scds: computational annotation of doublets in single-cell RNA sequencing data. *Bioinformatics*, 36(4):1150–1158, 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz698. URL http://dx.doi.org/10.1093/bioinformatics/btz698.

[35] Chenyu Li et al. Nephrobase Cell+: Multimodal Single-Cell Foundation Model for Decoding Kidney Biology. October 2025. doi: 10.1101/2025.09.30.679471. URL http://dx.doi.org/10.1101/2025.09.30.679471.

[36] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023. ISSN 2666-3899. doi: 10.1016/j.patter.2023.100804. URL http://dx.doi.org/10.1016/j.patter.2023.100804.

[37] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks, 2017. URL https://arxiv.org/abs/1706.04599.

[38] Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, February 2021. ISSN 2168-8966. doi: 10.1214/20-aos1965. URL http://dx.doi.org/10.1214/20-AOS1965.

[39] Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident Learning: Estimating Uncertainty in Dataset Labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, April 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12125. URL http://dx.doi.org/10.1613/jair.1.12125.

[40] Davin Lee, Gaeun Byeon, Seojin Chung, Dongmin Shin, Jongseo Park, Ingyeon Koh, and Joon-Yong An. AnnQ: reference-based quantification of cellular abnormality at single-cell resolution. *Briefings in Bioinformatics*, 27(3), May 2026. ISSN 1477-4054. doi: 10.1093/bib/bbag278. URL http://dx.doi.org/10.1093/bib/bbag278.

[41] Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21 (8):1481–1491, 2024. ISSN 1548-7105. doi: 10.1038/s41592-024-02305-7. URL http://dx.doi.org/10.1038/s41592-024-02305-7.

[42] Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorcan, Stephen R. Quake, and Jure Leskovec. Universal Cell Embeddings: A Foundation Model for Cell Biology. November 2023. doi: 10.1101/2023.11.28.568918. URL http://dx.doi.org/10.1101/2023.11.28.568918.

[43] Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. CellPLM: Pre-training of Cell Language Model Beyond Single Cells. October 2023. doi: 10.1101/2023.10.03.560734. URL http://dx.doi.org/10.1101/2023.10.03.560734.

[44] Xiaoyang Chen, Keyi Li, Xuejian Cui, Zian Wang, Qun Jiang, Jiacheng Lin, Zhen Li, Zijing Gao, Hairong Lv, and Rui Jiang. EpiAgent: foundation model for single-cell epigenomics. *Nature Methods*, 22(11): 2316–2327, 2025. ISSN 1548-7105. doi: 10.1038/s41592-025-02822-z. URL http://dx.doi.org/10.1038/s41592-025-02822-z.

[45] Achim Hekler, Lukas Kuhn, and Florian Buettner. Beyond Overconfidence: Foundation Models Redefine Calibration in Deep Neural Networks, 2025. URL https://arxiv.org/abs/2506.09593.

[46] Jan Engelmann, Leon Hetzel, Giovanni Palla, Lisa Sikkema, Malte Luecken, and Fabian Theis. Uncertainty Quantification for Atlas-Level Cell Type Transfer, 2022. URL https://arxiv.org/abs/2211.03793.

[47] Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, 2021. URL https://arxiv.org/abs/2107.07511.

[48] Tiffany Ding, Anastasios N. Angelopoulos, Stephen Bates, Michael I. Jordan, and Ryan J. Tibshirani. Class-Conditional Conformal Prediction with Many Classes, 2023. URL https://arxiv.org/abs/2306.09335.

[49] Robin Dunn, Larry Wasserman, and Aaditya Ramdas. Distribution-Free Prediction Sets for Two-Layer Hierarchical Models, 2018. URL https://arxiv.org/abs/1809.07441.

[50] Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candes, and Aaditya Ramdas. Conformal Prediction Under Covariate Shift, 2019. URL https://arxiv.org/abs/1904.06019.

[51] Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. 2021. doi: 10.48550/ARXIV.2103.14749. URL https://arxiv.org/abs/2103.14749.

[52] Can Ergen, Galen Xing, Chenling Xu, Martin Kim, Michael Jayasuriya, Erin McGeever, Angela Oliveira Pisco, Aaron Streets, and Nir Yosef. Consensus prediction of cell type labels in single-cell data with popV. *Nature Genetics*, 56(12):2731–2738, November 2024. ISSN 1546-1718. doi: 10.1038/s41588-024-01993-3. URL http://dx.doi.org/10.1038/s41588-024-01993-3.

[53] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, 2017. URL https://arxiv.org/abs/1703.04977.

[54] Lisa Sikkema et al. An integrated cell atlas of the lung in health and disease. *Nature Medicine*, 29(6):1563–1577, 2023. ISSN 1546-170X. doi: 10.1038/s41591-023-02327-2. URL http://dx.doi.org/10.1038/s41591-023-02327-2.

[55] Yuhei Kirita, Haojia Wu, Kohei Uchimura, Parker C. Wilson, and Benjamin D. Humphreys. Cell profiling of mouse acute kidney injury reveals conserved cellular responses to injury. *Proceedings of the National Academy of Sciences*, 117(27):15874–15883, 2020. ISSN 1091-6490. doi: 10.1073/pnas.2005477117. URL http://dx.doi.org/10.1073/pnas.2005477117.

[56] Louisa M. S. Gerhardt, Jing Liu, Kari Koppitch, Pietro E. Cippà, and Andrew P. McMahon. Single-nuclear transcriptomics reveals diversity of proximal tubule cell states in a dynamic response to acute kidney injury. *Proceedings of the National Academy of Sciences*, 118(27), 2021. ISSN 1091-6490. doi: 10.1073/pnas.2026684118. URL http://dx.doi.org/10.1073/pnas.2026684118.

[57] Kaichen Xu, Yan Lu, Suyang Hou, Kainan Liu, Yihang Du, Mengqian Huang, Hao Feng, Hao Wu, and Xiaobo Sun. Detecting anomalous anatomic regions in spatial transcriptomics with STANDS. *Nature Communications*, 15(1), September 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-52445-9. URL http://dx.doi.org/10.1038/s41467-024-52445-9.

[58] Jeffrey M. Granja, M. Ryan Corces, Sarah E. Pierce, S. Tansu Bagdatli, Hani Choudhry, Howard Y. Chang, and William J. Greenleaf. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nature Genetics*, 53(3):403–411, February 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00790-6. URL http://dx.doi.org/10.1038/s41588-021-00790-6.

[59] Asa Thibodeau et al. AMULET: a novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data. *Genome Biology*, 22(1), 2021. ISSN 1474-760X. doi: 10.1186/s13059-021-02469-x. URL http://dx.doi.org/10.1186/s13059-021-02469-x.

Appendix

A Injection recipes

This appendix gives, for each artifact, the recipe with its sources and the practical difficulties; Table 9 at the end compares the three. Statements marked "our reasoning" are not taken from a cited source.

The recipes below name their sources, which we checked against the papers and code on 31 October 2026. Parameter ranges, and whether injected artifacts look like real ones, remain assumptions to be tested on real answers.

Doublets: mostly summed, some library sizes shrunk. Published tools differ. Scrublet sums the raw counts of two random cells ("Doublets are simulated by adding the unnormalized counts"), simulating twice as many doublets as cells [9]. DoubletFinder averages two cells' counts; a detector trained this way recovered 73% and 64% of the doublets labelled by demuxlet and Cell Hashing [10]. scDblFinder by default sums 75% of pairs and, for the other 25%, halves the sum and Poisson-resamples it; it pairs only across clusters, in proportion to cluster size, after keeping cells within the 5–95th library-size percentiles [11]. The reason to shrink some library sizes is that in genotype-labelled real data, median doublet library sizes were "systematically smaller than the sum" of their two cell types, and summing and size re-weighting were both "clearly superior to averaging" [11]. We follow scDblFinder: pair across cell types in proportion to expected frequency, sum 75%, halve and resample 25%, and simulate 0.8–2 times as many doublets as cells. Same-type doublets are "virtually unidentifiable" [11], so they get their own label and stay out of the main metrics. Methods score higher on simulated than on real doublets, with a best mean AUPRC of only 0.537 on real labels [18]; the genotype doublets of Hoeft et al. [4] are the test. Simulating doublets is easy; doing it well needs three further rules (our reasoning). Pairs are drawn only within one sample, because real doublets form only within one channel. The base data already contain real doublets, so a first pass with an existing detector should remove the most obvious ones, or the paper must state that injected pairs may occasionally include them. Cells used to build doublets and cells used for testing come from different donors. Since doublet simulation is standard practice, the contribution here is not the recipe but whether boosting narrows the gap between simulated and real doublets, and whether models trained on several artifacts separate them from disease.

Ambient RNA: a background profile mixed in at a set fraction. Removal models can be used in reverse. SoupX writes observed counts as the cell's own counts plus a contamination fraction times a background profile estimated from empty droplets, and "assumes a constant contamination fraction within a channel" [22]. DecontX draws each cell's contamination fraction "from a global beta distribution" and by default takes the contamination profile as "a weighted combination of all other cell population distributions", or empty droplets if given [23]; its package includes a simulation that keeps total counts fixed and replaces a binomially drawn share of counts, though on fully synthetic matrices, so we reuse the recipe rather than the function. CellBender treats background as noise added to each cell, with the profile learned from cell-free droplets [24], and an ambient-removal benchmark simulates per-cell fractions with ambient counts drawn from other cell types [5]. For calibration, Janssen et al. [3] measured by genotype an average of 3–35% background per cell in mouse kidney, far more in single nuclei from frozen tissue than in single cells (35% vs 11%, 17% vs 3%), with the nuclear contamination profile closer to empty droplets (Spearman $\rho \approx 0.85$) than to the cells themselves (≈ 0.50). We draw a per-cell fraction from a Beta distribution with its mean swept over 3–35%, try both replacement at fixed depth and background added on top, and store native and injected counts separately. The main human datasets [1, 19] have no empty droplets; the profile can then be approximated from the other cell populations or the sum over all cells, but since nuclear background resembles empty droplets, better profiles come from Janssen et al. [3] (mouse) or a re-run of Hoeft et al. [4] (human kidney).

A background without empty droplets. The background profile can be synthesised from the cells themselves, as a weighted mixture of cell-type mean profiles. This has precedent: the CellBender simulation code builds its ambient profile as a weighted average of expression, each cell type weighted by its number of cells and mean UMI count (`simulate.py`); and the ambient-removal benchmark samples contaminating counts from the other cell types' distributions [5]. We suggest drawing the mixture weights at random (e.g. from a Dirichlet distribution) so that models see many possible backgrounds rather than one; since injected artifacts are used for training only, this randomisation is allowed. Janssen's data contain both cells and empty droplets, so they can measure how far a synthesised background is from the real one, and how much the difference changes the trained models; if the gap is small, human data need no empty droplets.

Difficulties. (i) *Realism.* Real soup comes mainly from fragile cells and, in nuclei, carries cytoplasmic mRNA that nuclear means lack [3], so a model may learn the synthetic background rather than contamination itself; randomised weights reduce but do not remove this risk. (ii) *A continuous artifact.* Unlike a doublet, which a droplet either is or is not, every cell carries some ambient RNA (3–35% per cell on average in mouse kidney, varying across samples and

cells [3]), so the level that counts as an artifact is our choice. We treat it either as regression on the injected fraction or in reported bands (low, medium, high). If each droplet takes up a similar amount of soup, cells with little RNA have a higher fraction, which blurs ambient RNA with low-quality cells (our reasoning). (iii) *The base is not clean.* The base data already contain real ambient RNA, so labels cover only the injected part; decontaminating the base first with an existing tool trades one approximation for another, and the paper must say which was done. (iv) *Overlap with biology.* Contaminated cells express another cell type's markers, and so can real transitional or injured cells; this is the distinction the models must learn.

Low-quality cells: approximate only, and different for cells and nuclei. In microscopy-labelled broken cells, "cytoplasmic RNA will be lost, but RNAs enclosed in the mitochondria will be retained", and mtDNA-encoded genes are only "marginally upregulated" [25] (mouse ES cells on Fluidigm C1). For single cells, the only ready simulator is an R package that removes a Beta-distributed fraction (default 0.1–0.8) of cytoplasmic gene counts by abundance-weighted sampling, keeping mitochondrial genes and nuclear lncRNAs [12]; it has no paper, no peer review and no validation against experimentally labelled damaged cells, and other QC tools detect but do not simulate. Our main data are single-nucleus, and nuclei have no cytoplasm: in kidney single-nucleus data mitochondrial transcripts are "not expressed in nuclei" and are attributed to mitochondria attached to the nuclear membrane [26], and in contaminated nuclei the mitochondrial share rises and the unspliced share falls, "as mitochondria are only present in the cell cytoplasm" [27]. Low-quality nuclei are therefore made by *adding* cytoplasmic material (mitochondrial counts and spliced reads) rather than removing it, which brings them close to ambient injection; the design must draw the line between the two. Kidney proximal tubule naturally has a high mitochondrial share, and a fixed 10% cut-off removes real kidney parenchymal cells [28], so PT is scored separately. Degraded nuclei may also have fewer counts overall, which can be imitated by random down-sampling (our reasoning). Low-quality cells are the hardest of the three: there is no validated simulator; for nuclei the recipe is almost the same as ambient injection, so the two must be separated by design or the leave-one-artifact-out test is not clean; and the degenerative state of Lake et al. [1] was itself partly defined by few genes and many mitochondrial transcripts, so testing against it is partly circular (the known trap in the evaluation section). Above all, *no public kidney dataset gives real answers for low-quality cells*. The recipe can therefore only be checked indirectly: by comparing injected cells with the cells that standard quality control flags in the data of Lake et al. [1], and by leave-one-artifact-out (train on doublets and ambient RNA, test on low-quality cells).

Table 9: The three artifacts compared: recipe, maturity, main difficulty and real answer.

	Doublet	Ambient RNA	Low-quality cell
Recipe	sum two cells of different types; shrink a quarter of library sizes	mix a background profile in at a per-cell fraction	single cells: remove cytoplasmic RNA; single nuclei: add cytoplasmic material
Maturity	mature	removal models used in reverse	approximate only
What counts as an artifact	clear: a droplet is a doublet or not	continuous: the threshold is our choice	continuous, and overlaps the degenerative state
Main difficulty	simulated doublets are easier than real ones	realism of the background; the base already contains ambient RNA	no validated simulator; confounded with ambient RNA and with disease
Real human-kidney answer	Hoeft et al. [4] genotype doublets	Hoeft et al. [4], after a re-run from FASTQ	none
Order	first	second	third

B Foundation-model candidates

Pretraining from scratch is unrealistic: the first kidney-specific multimodal foundation model, Nephrobase Cell+ [35], used about 39.5 million cells, four species and five assays, with up to one billion parameters, and as of 28 October 2026 we found no public code or weights. Table 10 lists the open candidates checked on that date. We found no model pretrained jointly on RNA and ATAC with public weights, so stage 2 may pair an RNA model with an ATAC model.

C Uncertain cells: out-of-fold scores

Models can also flag cells they are unsure about, and all models are scored the same way (suggested, not decided). (i) *Donor-grouped K-fold cross-validation, out-of-fold scores only.* Each cell is scored by the fold model that never saw it, and folds are split by donor, since the same person in training and test is a recognised form of leakage [36]. Among doublet tools, scds already scores cells by 5-fold cross-validation, but with random cell folds [34], while scDblFinder

Table 10: Open foundation-model candidates (checked 28 October 2026).

Model	Licence and weights	Modality	Role
Geneformer V2 [31]	Hugging Face, Apache-2.0; 104M and 316M parameters	RNA	first choice for fine-tuning
scGPT [32]	code MIT; weights on Google Drive with no separate weight licence stated; includes a checkpoint pretrained on 814 thousand kidney cells	RNA	fine-tuning; its multi-omic recipe trains the non-RNA parts from scratch
scFoundation [41]	weights for non-commercial research only	RNA	comparison
UCE [42], CellPLM [43]	public weights	RNA	frozen-embedding audits
EpiAgent [44]	MIT, with fine-tuning code	scATAC	stage 2 candidate

uses cross-validation only to choose the number of boosting rounds and scores the cells it was trained on (our reading of its code; 11). (ii) *Calibrate, then threshold.* Neural networks are often overconfident and temperature scaling is a simple fix [37]; foundation models can instead be underconfident, with post-hoc calibration less reliable under strong shift [45], and single-cell annotation models also lack calibration [46]. (iii) *Conformal thresholds.* CV+ extends the jackknife+ to K folds with a coverage guarantee [38]; it is stated for regression, so applying it to classification scores is our adaptation. Per-cell-type guarantees come from calibrating within each class [47], which is weak when a class has few cells [48]; cells from one person are not exchangeable [49], and in a new laboratory exchangeability fails altogether: single-cell conformal annotation lost its coverage guarantee on disease data [7], weighted conformal is one possible remedy [50], and T4 tests exactly this. (iv) *Uncertain cells go to experts but are never used as labels.* Confident learning finds suspect labels from cross-validated scores [39], yet only 51% of its flagged candidates were confirmed by people [51]; popV uses agreement across methods to reduce manual inspection [52]. Uncertainty from noise in the data differs from uncertainty about what the model has not seen [53], and a score alone cannot say whether a cell is an artifact, a transitional state or early disease; AnnQ notes that such cells are "typically treated as technical artifacts and filtered out" although they may be the meaningful deviations [40]. We therefore report, among the cells sent for review, the shares experts call artifact, disease or unclear, and the trade-off between expert workload and error rate. To our knowledge (October 2026), donor-grouped out-of-fold scores have not been used in single-cell annotation or quality control, nor has cross-validated or conformal uncertainty been used to separate artifact from disease.

D Clinical experts for the disease answer

The disease labels are expert judgement, the weakest link. Clinical experts (nephrologists, renal pathologists) can strengthen them in two ways. **(A) Re-annotate clusters.** Experts see each altered cluster's marker genes (e.g. HAVCR1, LCN2), QC distributions and donor conditions, and call it likely injury, likely low quality or unsure, with two or three independent annotators and agreement reported. The Human Lung Cell Atlas re-annotated per cluster with six experts, most of whom agreed or differed only in granularity, without reporting κ [54]; kidney injury states have independent anchors in mouse, such as Havcr1 checked by immunofluorescence [55] and genetically labelled injured PT cells [56]. This involves a few hundred clusters, not hundreds of thousands of cells; clinicians may need a short calibration set, and one single-cell analyst should be among the annotators. We found no single-cell study reporting κ between human annotators, so the agreement measured here is itself a result. **(B) Annotate histology.** Pathologists outline injured regions on H&E slides paired with spatial transcriptomics, and anomaly scores are checked spot by spot. The Visium data of Lake et al. [1] (GEO GSE183456) contain 23 sections (6 reference, 11 diabetic kidney disease, 6 acute kidney injury), each with a full-resolution H&E image; Abedini et al. [20] provide 13 H&E images. We found no public pathologist region annotation, so it would have to be made. A Visium spot is 55 µm wide and holds several cells, so labels are per spot; only three Visium participants also appear in the single-nucleus data (our count), so this would be a separate spatial task rather than a check on T1.

E Future plan: a new model?

The first round uses existing models only (not decided). The main claim is whether the pipeline helps, and changing the model at the same time would blur the credit; the existing models already include deep models; and a new model need not win, since no unsupervised method was better overall on tabular data [29] and foundation models zero-shot were often behind simple methods [15]. After the first round: if the artifact discriminator clearly improves every omics model, no new model is needed; if it recognises injected artifacts but fails on real answers, a model that learns the anomaly score and the artifact call together is worth building; if only fine-tuned foundation models improve, the contribution is

more likely a way to fine-tune omics models on artifacts. The candidate is a single-cell analogue of DRAEM [16]: a reconstruction branch trained on healthy cells, whose error is the anomaly score, and a discriminative branch trained on injected artifacts that outputs their type, trained together. Searching up to 31 October 2026 we found no such model in single-cell or spatial omics; the closest, STANDS, trains a reconstruction GAN on a normal reference in spatial transcriptomics but does not train on synthetic anomalies [57]. An idea specific to this topic, not yet checked against the literature, is to penalise reliance on gene count and mitochondrial share so the model does not learn only "poor quality"; this may also suppress real injury signal.

F Stage 2: ATAC

RNA comes first; ATAC is added once the RNA stage works (suggested, not decided). ATAC has injection precedents: ArchR generates synthetic doublets and finds their neighbours [58]; AMULET simulates doublets by adding two nuclei's read counts [59]; ambisim controls ambient molecule proportions in joint snRNA/snATAC data [13]. Low-quality nuclei are judged by TSS enrichment, fragment counts and similar metrics, with thresholds tuned per dataset.