Making Single-Cell Data Distillation Auditable: Traceable Real-Cell Coresets via Discrete Min–Max Selection

Yaodi Luo[2*], **Peize He**[2*], **Bowen Han**[2], **Lingbei Meng**[1,2†]

[1]The Chinese University of Hong Kong, Shenzhen
[2]Shenzhen Loop Area Institute Shenzhen

2023300904024@std.uestc.edu.cn 2023300904027@std.uestc.edu.cn
hanbowen@student.usm.my 225085001@link.cuhk.edu.cn

Abstract

Single-cell datasets are increasingly costly to store, audit, and reuse for model training. Dimensionality reduction and dataset distillation can reduce this burden, but conventional distillation methods often produce synthetic expression profiles that cannot be traced to an assayed cell. We formulate traceable single-cell data distillation as retaining original cell identifiers and gene symbols under fixed cell and gene budgets. The resulting training subset remains connected to measured counts, labels, and assay metadata, so unexpected predictions can be checked against their source data. We propose two real-cell selectors. Fixed-CF uses static characteristic-function matching. Minmax-CF solves an entropy-regularized discrete min–max problem that upweights poorly preserved directions and adds only observed cells. Across donor-, technology-, and perturbation-level shifts on three datasets, Minmax-CF retains 96.52% of Full balanced accuracy on MS, approximately matches Full on average on hPancreas with a median $2.55\times$ GPU speedup in the all-gene setting, and obtains the lowest pathway error among compressed methods on Norman. Performance remains weaker for rare states, some technology shifts, unseen perturbation components, and settings where fidelity is weakly associated with downstream utility. Because the selected IDs refer to measured cells, these cases can be investigated by inspecting the corresponding training support, labels, and assay metadata. Minmax-CF consistently reduces worst-direction discrepancy, while downstream utility and cost vary across datasets and tasks.

Introduction

Single-cell atlases now span more donors, tissues, technologies, conditions, and perturbations, but the resulting matrices are costly to reuse. Model comparison, hyperparameter search, annotation updates, and cross-site validation repeatedly revisit the same cells. PCA and nonlinear embeddings such as t-SNE and UMAP provide low-dimensional representations (van der Maaten and Hinton 2008; McInnes, Healy, and Melville 2018), while geometric sketching and metacells summarize cellular state space (Hie et al. 2019; Persad et al. 2023). These outputs are valuable, but they do not by themselves define a compact reusable training dataset in which every row remains an assayed cell. We study *traceable single-cell data distillation*: retaining a compact set of original cell identifiers $\times$ original gene symbols and measuring distribution fidelity, downstream utility, and computation.

Row identity matters because a biological result may need to be interpreted against its donor, assay, condition, and original count profile. Unlike general AI settings, where synthetic support may suffice when predictive behavior is preserved, our downstream predictors train on a subset of measured cells. When a model fails or yields a surprising result for a rare state, donor, technology, or perturbation, a retained ID permits direct return to its counts, label, quality-control fields, and provenance. Researchers can test whether the failure reflects absent training support, selector omission, misannotation, or technical bias, revise the data or selection process, and potentially recover an overlooked biological signal. Synthetic datasets and their generators can still be audited, but a generated profile has no unique assayed-cell source and cannot support the same direct row-level recheck. This traceability enables a prediction–audit–revision loop; it does not guarantee a causal diagnosis or a better downstream model.

Dataset distillation commonly learns synthetic examples that reproduce training behavior through gradients, distributions, or trajectories (Wang et al. 2018; Zhao, Mopuri, and Bilen 2021; Zhao and Bilen 2023; Cazenavette et al. 2022). In single-cell analysis, scDD is the closest precedent (Yu et al. 2025): it transfers information from the original data and a foundation model into optimized latent codes, then generates synthetic expression profiles with conditional diffusion. Its cross-architecture experiments establish compact synthetic scRNA-seq training data as a credible alternative to the full matrix. They also expose the support choice that motivates our formulation. scDD optimizes generated profiles for compact utility and transfer; we optimize which measured cells to retain for fidelity and traceability. Because its datasets, budgets, splits, preprocessing, and evaluators differ from ours, this is a formulation-level comparison rather than an unmatched numerical claim.

Restricting the support to real cells turns continuous synthesis into finite selection, but selection remains nontrivial. Stratified Random sampling can be strong when its strata capture dominant variation, whereas sketching and medoids emphasize local geometric coverage. Neither local coverage nor an average discrepancy guarantees preservation of a rare state or condition-associated direction with large error. We

*These authors contributed equally.
†Corresponding author.

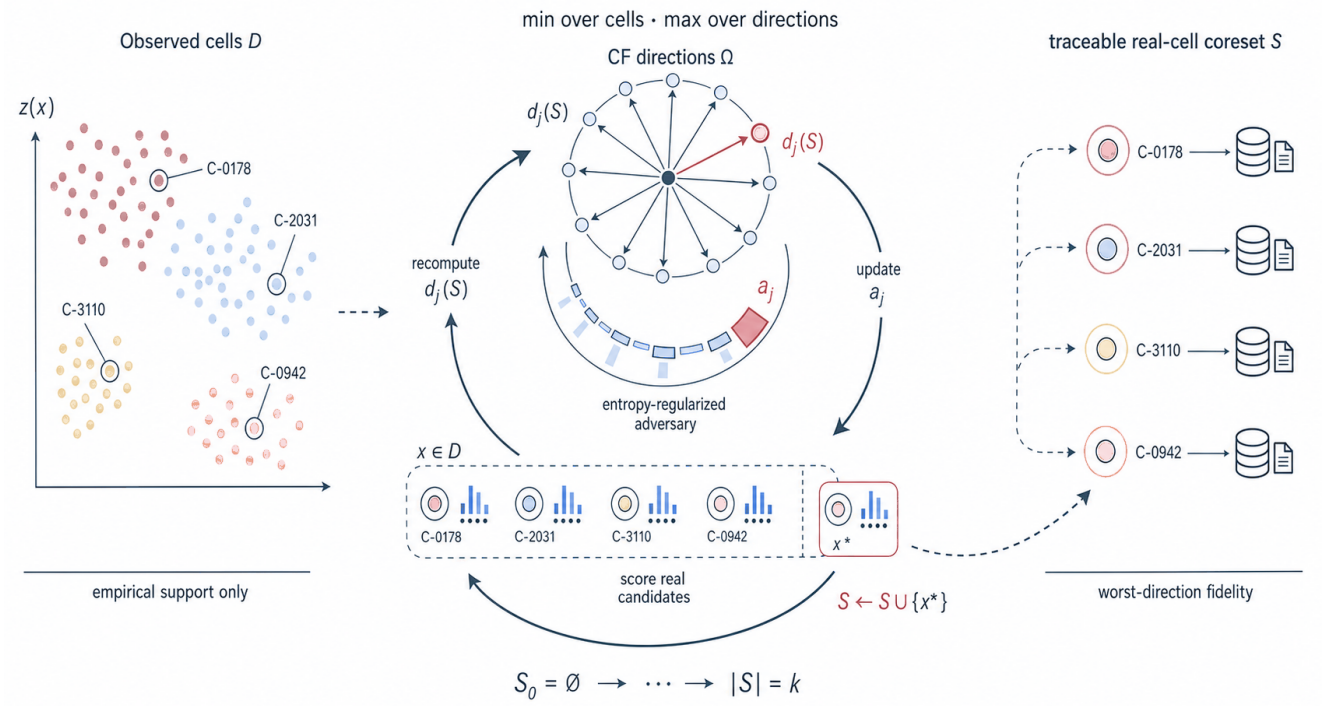

Figure 1: Minmax-CF selects a traceable coreset entirely from observed cells. Within each biological stratum, the entropy-regularized adversary emphasizes CF directions with large current discrepancy; the selector scores measured candidates under those weights, adds the best real cell, and recomputes the discrepancies until the stratum budget is reached. The per-stratum selections are concatenated, and every selected ID retrieves its original expression profile and provenance.

propose two CF-based real-cell selectors. Our base method, **Fixed-CF**, greedily matches a fixed pool of characteristic-function directions using only measured candidates. Our core method, **Minmax-CF**, adds an entropy-regularized adversary whose weights evolve with the subset: it emphasizes the currently worst-preserved directions, and each greedy step adds the observed cell that best repairs their weighted discrepancy. This discrete min–max construction adapts adversarial characteristic-function matching (Wang et al. 2025) without allowing the support to leave an assayed cell. Fixed-CF isolates the effect of static CF matching; Minmax-CF tests whether dynamic worst-direction reweighting improves the stated fidelity objective.

We evaluate the distilled data rather than only the selector objective. Preprocessing and selection are fit inside each training split, with donors, sequencing technologies, or perturbation groups held out for testing. Under matched budgets, we compare Full data, stratified Random, PCA k-medoids, and our Fixed-CF and Minmax-CF selectors. Across MS, hPancreas, and Norman, Minmax-CF reduces matched soft and worst-direction CF discrepancies by approximately 42%–52%. Downstream utility is more conditional: the real-cell coresets retain much of Full utility and can reduce repeated training cost, but PCA or Random is equally good or better in some regimes. Lower worst-direction discrepancy therefore validates the declared fidelity objective, not universal downstream superiority.

Our contributions are:

- a traceable single-cell data-distillation formulation that returns original cell IDs × original gene symbols and enables prediction failures to be rechecked against measured profiles and provenance;
- Two observed-cell selectors: Fixed-CF for static characteristic-function matching and the core Minmax-CF method for entropy-regularized, worst-direction-aware discrete selection.
- A train-only, leakage-resistant evaluation across donor, technology, and perturbation shifts.
- A fidelity–utility–cost analysis that separates reproducible direct-objective gains from failure regimes in which simple selectors remain equally good or better.

Related Work

Dataset distillation and distribution matching. Dataset distillation replaces a large training set with a compact object that preserves training value. Early work optimized synthetic examples through the training process (Wang et al. 2018); later methods matched gradients (Zhao, Mopuri, and Bilen 2021), expert trajectories (Cazenavette et al. 2022), or feature distributions (Zhao and Bilen 2023). Neural Characteristic Function Matching (NCFM) makes the distribution discrepancy adversarial by learning frequency directions that distinguish real and synthetic samples (Wang et al. 2025).

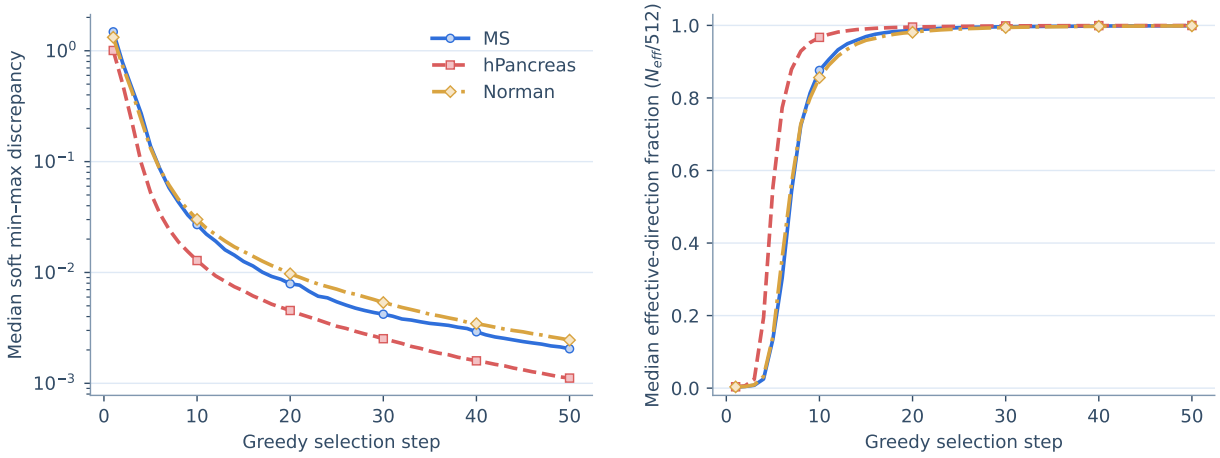

Figure 2: Minmax-CF optimization dynamics over 50 greedy steps. Lines are medians across the available biological-group traces for each dataset. Left: entropy-smoothed min–max discrepancy on a logarithmic scale. Right: the effective number of adversarially weighted directions divided by $M = 512$. Step 1 is measured after the first adversarial update, not at the uniform initialization. The increasing fraction shows that the weights become less concentrated as the largest discrepancies are repaired.

These methods differ in what they preserve, but most optimize a continuous synthetic support. Minmax-CF retains the worst-direction distribution-matching principle while restricting the output to a finite subset of observed cells.

Synthetic and real-cell summaries for single-cell data. scDD is the closest single-cell distillation method (Yu et al. 2025). It optimizes latent codes and maps them through a conditional diffusion generator to produce synthetic expression profiles, using single-cell foundation-model representations and matching losses to guide the process. Its cross-architecture experiments test whether the generated dataset transfers beyond the model used for distillation. Our comparison is formulation-level rather than numerical: scDD's datasets, random split, evaluator, preprocessing, and per-class budgets differ from our donor-, technology-, and perturbation-OOD protocols. scDD allows synthetic support; Minmax-CF asks what can be retained when every output row must remain an assayed cell with recoverable provenance.

Real-cell summarization provides the other relevant comparison. Stratified Random sampling is transparent but can miss within-stratum diversity. Geometric sketching samples observed cells to cover transcriptional state space and improve representation of rare populations (Hie et al. 2019); PCA k-medoids similarly targets local geometric coverage. SEACells instead aggregates neighborhoods into archetypal metacells (Persad et al. 2023), which can stabilize noisy states but no longer yields a one-to-one assayed-row output. These approaches motivate strong coverage baselines, but they optimize a different criterion: local representativeness does not guarantee small error in every global distribution projection. We therefore compare Random and PCA medoids directly and report geometric and characteristic-function fidelity separately.

Evaluation under biological shifts. A random cell split can place cells from the same donor, platform, or perturbation condition in both training and test data. Leakage can also occur when feature selection, scaling, PCA, or frequency directions are fit before the split. We instead fit preprocessing and selection within each training fold and hold out donors, sequencing technologies, or perturbation groups. This distinction is especially important for combinatorial perturbations. Norman Perturb-seq exposes structured genetic responses (Norman et al. 2019), and GEARS targets unseen multigene perturbations (Roohani, Huang, and Leskovec 2023); however, recent evaluations show that simple additive or linear baselines remain competitive in important regimes (Ahlmann-Eltze, Huber, and Anders 2025). We consequently separate seen and unseen perturbation components and distinguish direct fidelity, local geometry, and downstream prediction rather than treating any one as a universal surrogate.

Traceable Single-Cell Data Distillation

Output contract. Let $D^{\mathrm{tr}} = \{(x_i, c_i, m_i)\}_{i=1}^{n}$ be a training split, where x_i is the measured expression vector, c_i is a biological stratum (e.g., cell type–condition or perturbation group), and m_i contains provenance fields. We distill two finite index sets: original cell identifiers $S \subseteq [n]$ and original gene symbols $H \subseteq G$. The released training object is therefore the submatrix $X[S, H]$, together with the source identifiers needed to retrieve every entry. It contains neither synthetic expression profiles nor latent prototypes. Metadata m_i is used to form biological strata and to audit donor, study, platform, condition, and perturbation coverage; it is not supplied as a predictive feature to the selector or downstream model.

Information boundary. The train/validation/test partition is fixed before distillation. Normalization statistics, variable-gene scores, scaling, PCA coordinates, frequency pools, gene-panel ranks, and cell choices are fit using D^{tr} only. Val-

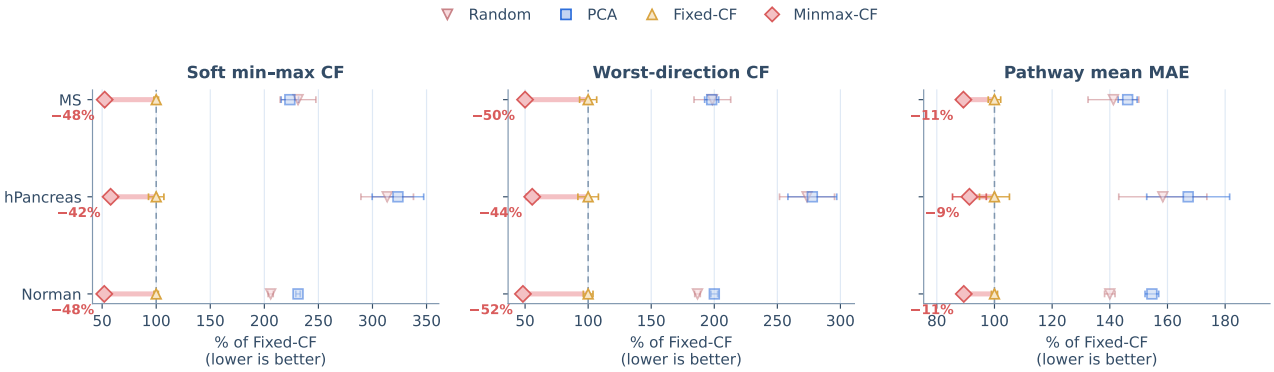

Figure 3: Normalized direct-objective evidence (lower is better). The first two panels use a shared 512-direction evaluation pool; the third reports the separate pathway-mean error diagnostic. The dashed line fixes Fixed-CF at 100%; horizontal segments show the change from Fixed-CF to Minmax-CF, while lighter Random and PCA points provide context. Points are means and error bars are standard deviations across the available seed or fold-seed units. The comparison measures distribution fidelity, not downstream superiority.

idation data can select a prespecified configuration; the held-out donor, technology, or perturbation group is untouched until final evaluation. This boundary is essential: a traceable subset can still leak if its representation or panel was learned globally.

Two budgets, one evaluation. Cell budget k_c and gene budget k_g are distinct axes. Our method contribution concerns S: Minmax-CF selects real cells within each biological stratum. The gene axis is controlled using all/common genes or a train-only 500-gene panel (HVG, random expressed genes, or a locked biological panel). We evaluate a distilled object by a vector rather than a single score:

$$\mathcal{E}(S, H) = \big(\text{fidelity},\ \text{OOD utility},\ \text{selection/training cost}\big).$$

The target is the fidelity–utility–cost Pareto frontier. In particular, lower distribution discrepancy is treated as a direct objective, not as proof of higher downstream accuracy.

Minmax-CF

Fixed average matching can hide a small set of poorly preserved projections. Inspired by the neural-CF min–max formulation for synthetic data (Wang et al. 2025), Minmax-CF changes both the feasible set and the optimizer: it allocates a finite adversary over characteristic-function (CF) directions and greedily repairs the largest current errors using only observed cells.

Worst-direction objective. Within one biological stratum, let D denote the available training cells and $z(x)$ their train-fitted representation. For a shared pool $\Omega = \{\omega_j\}_{j=1}^{M}$, define the empirical characteristic function

$$\Phi_A(\omega_j) = \frac{1}{|A|} \sum_{x \in A} \exp\big(\mathrm{i}\, \omega_j^\top z(x)\big)$$

and direction error $d_j(S) = |\Phi_S(\omega_j) - \Phi_D(\omega_j)|^2$. We seek a size-$k$ subset of original cells:

$$\min_{S \subseteq D,\, |S|=k}\ \max_{a \in \Delta_M} \sum_{j=1}^{M} a_j d_j(S) - \tau\, \mathrm{KL}(a \| u), \qquad (1)$$

where u is uniform over directions. Entropy regularization prevents the adversary from collapsing permanently onto one noisy projection while retaining sensitivity to worst directions. For a fixed S, the inner optimum is

$$a_j(S) = \frac{\exp(d_j(S)/\tau)}{\sum_{\ell=1}^{M} \exp(d_\ell(S)/\tau)}. \qquad (2)$$

Discrete real-cell herding. Equation 1 is combinatorial. We use a discrete greedy procedure that never leaves the empirical support:

1. For each biological stratum, initialize $S_0 = \varnothing$ and uniform frequency weights.
2. At step t, compute $d_j(S_t)$ and update a_j using Equation 2.
3. For each cell x in a capped training-only candidate pool, score $S_t \cup \{x\}$ by $\sum_j a_j d_j(S_t \cup \{x\})$.
4. Add the minimizing candidate's original cell ID and repeat until the stratum budget is met.
5. Concatenate the per-stratum IDs and attach their original gene symbols and provenance fields.

The fixed-CF baseline uses the same discrete construction but does not adapt an adversary to the current worst directions. PCA k-medoids instead targets local geometric coverage; these are different preservation criteria and neither dominates by definition.

Figure 1 summarizes the adaptive selection loop and the traceability of its output.

Locked configuration and cost. Before downstream evaluation, we fixed $M = 512$, $\tau = 0.05$, a candidate cap of 700 cells per stratum, and a budget of at most 50 cells per

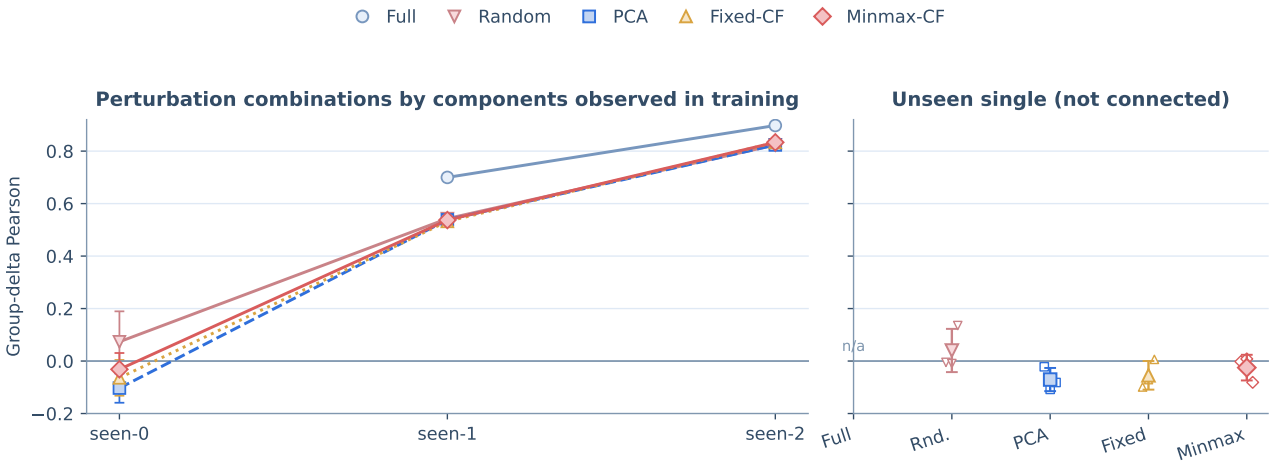

Figure 4: Norman group-delta Pearson correlation stratified by perturbation-component support. Full and all compressed selectors perform best when both components are observed and degrade when support is partial. Full correlations are undefined for seen-0 combinations and unseen singles because the additive predictor lacks train-derived component estimates; bars are shown only for compressed selectors in these regimes. Error bars are standard deviations across three selector seeds. The remaining Full–compressed differences in seen-1 and seen-2 are reported in the text.

Table 1: Datasets and leakage-resistant evaluation units.

Dataset	Cells	Genes	Held-out unit
MS	21,312	3,000	Individual/donor
hPancreas	16,382	19,093	Technology (9 folds)
Norman	91,205	5,045	Perturbation group

biological group. The greedy score requires more one-time work than random, PCA, or fixed-CF selection, but its output is an ordinary sparse slice of the original count matrix. It introduces no generator, external checkpoint, or inference-time selector. Minmax-CF optimizes worst-direction distribution preservation; it never observes downstream test labels or directly optimizes a classifier or perturbation loss.

Experimental Protocol

Datasets and shifts. Table 1 summarizes the three formal benchmarks. MS is an 18-class annotation task with donor/individual-disjoint train, validation, and test partitions. hPancreas uses nine leave-one-technology-out folds; every fold refits preprocessing and selection without its test technology. Norman uses the official perturbation split, with test combinations stratified by whether zero, one, or two component perturbations are observed in training, plus unseen single perturbations. IFNB is retained only as a historical joint cell–gene compression pilot: its study identity is confounded with stimulation, so it is not used for an OOD generalization claim.

Selectors and gene settings. We compare the uncompressed training split (Full), stratified Random, PCA k-medoids, Fixed-CF, and Minmax-CF under the same per-group budget. All compressed selectors preserve every nonempty training stratum. The formal multi-dataset evidence uses all/common genes and train-only HVG-500 where available. Random expressed genes and a locked 500-gene biological panel are controlled gene-axis baselines; they are emphasized only in the IFNB pilot because corresponding multi-seed OOD coverage is incomplete.

Downstream models and statistical units. Annotation uses logistic regression and a one-hidden-layer MLP with preprocessing fit on the current training subset. Norman uses a train-only additive-response baseline and evaluates condition-level expression deltas. MS reports three random seeds. hPancreas reports 9 technology folds $\times$ 3 seeds (27 fold-seed units) for PCA and Minmax-CF; where older Random or Fixed-CF runs contain only one seed, we label $n = 9$ rather than pooling unequal evidence silently. Technology-cluster bootstrap resamples the nine technologies with 10,000 replicates. Norman reports three seeds. We do not treat cells as independent replicates for confidence intervals.

Metrics. For annotation, we report balanced accuracy (BA), macro-F1, recall over rare classes defined from training counts, and retention relative to a matched Full run. Perturbation utility is group-delta Pearson correlation and pathway-delta mean absolute error (MAE), with pathway groups derived from Gene Ontology and Reactome resources (Gene Ontology Consortium 2023; Milacic et al. 2024). Direct fidelity metrics are entropy-smoothed min–max CF discrepancy, worst-frequency CF discrepancy, pathway-mean MAE, and nearest-PCA distance. The last metric favors local geometric coverage, whereas the CF metrics evaluate distribution projections. All selector comparisons use the same held-out direction pool.

Compute accounting. We separate one-time selector wall time from downstream GPU training time and peak allocated

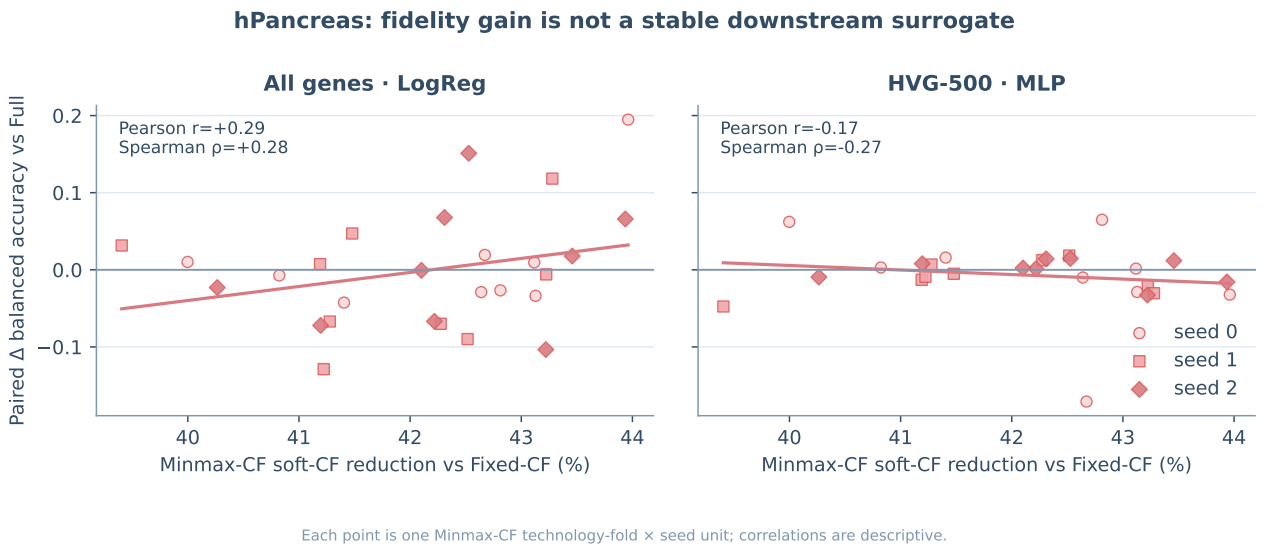

Figure 5: Relationship between Minmax-CF's direct fidelity gain and downstream hPancreas utility. Each point is one matched technology-fold × seed unit: the horizontal coordinate is Minmax-CF's soft-CF reduction relative to Fixed-CF, and the vertical coordinate is its paired ΔBA relative to Full.

memory. Speedup is computed against the matched Full configuration on the same fold, model, and feature space. Since selection can be reused across models and hyperparameter searches, we report its natural wall time separately and discuss amortization rather than subtracting it from every downstream run.

Results

Worst-direction discrepancy. Figure 3 evaluates Fixed-CF and Minmax-CF on the same held-out pool of 512 directions. From Fixed-CF to Minmax-CF, soft/worst-direction CF discrepancy falls by $47.6\%/50.0\%$ on MS, $42.2\%/44.3\%$ on hPancreas, and $48.0\%/51.7\%$ on Norman. Pathway-mean MAE simultaneously falls by 10.8%, 8.7%, and 10.6%, respectively. These consistent reductions show that adversarial reweighting lowers the target discrepancy across datasets. PCA retains the lowest mean nearest-PCA distance on each dataset, indicating that local geometric coverage and worst-projection matching measure different properties.

Adversarial-weight dynamics. Figure 2 reports group-level optimization traces. Immediately after the first adversarial update, the effective weight is concentrated on only a few directions. As greedy selection reduces the soft min–max discrepancy, the effective direction fraction approaches one, indicating that errors across the finite pool have become more even. The trace records how the weighting adapts during selection at the group level.

MS donor-OOD annotation. Figure 6 summarizes the MS annotation evidence. With all-gene logistic regression over three seeds, Full obtains 0.7729 BA. Minmax-CF obtains 0.7459, or 96.52% retention, compared with 0.7444 for PCA, 0.7431 for Fixed-CF, and 0.7369 for Random (Table 2). Minmax-CF and PCA differ by only 0.0015 BA. Minmax-CF rare recall is 0.7039, above Full (0.6143) but below Fixed-CF (0.7206) and the prespecified 0.90 target. Average BA retention is high, but rare recall remains below the target. Because the coreset retains original IDs, a follow-up audit can inspect which rare cells and donors were selected. The current aggregate recall cannot identify individual omissions.

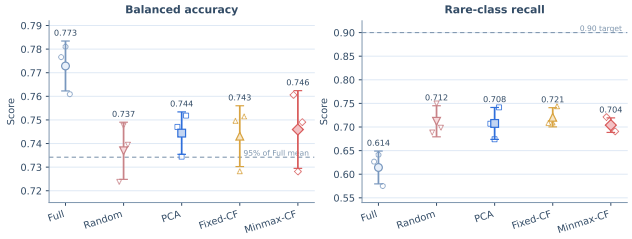

Figure 6: MS donor-OOD annotation across three seeds. Hollow points are seed observations; filled points and intervals show means and standard deviations for balanced accuracy and rare-class recall. Dashed lines mark 95% of the Full mean BA and the prespecified 0.90 rare-recall target.

hPancreas technology-OOD annotation. Figure 7 shows the matched fold-seed differences. For all-gene logistic regression, Minmax-CF has mean paired $\Delta\text{BA} = -0.0010$ against matched Full, with a technology-cluster bootstrap interval $[-0.0411, 0.0455]$ and median GPU speedup $2.55\times$. For HVG-500 plus MLP, its ΔBA is -0.0069 with interval $[-0.0254, 0.0067]$, rare recall 0.967, and speedup $2.09\times$. Both intervals include zero. After averaging the three seeds within each held-out fold, the worst Minmax-CF fold has $\Delta\text{BA} = -0.0887$ (-8.87 percentage points) for all-gene lo-

Table 2: OOD utility. MS and Norman are three-seed means; hPancreas n counts fold-seed units. Bold marks the best compressed method; Full is uncompressed.

Method	MS BA ↑	Norman r ↑	Path. MAE ↓
Full	0.7729	0.7572	0.03317
Random	0.7369	**0.3897**	0.03530
PCA	0.7444	0.3385	0.03927
Fixed-CF	0.7431	0.3455	0.03516
Minmax-CF	**0.7459**	0.3597	**0.03465**

hPancreas all-gene LogReg	n	ΔBA	GPU speedup
Full	27	0	1.00×
Random	9	+0.0046	2.27×
PCA	27	−0.0003	2.84×
Fixed-CF	9	+0.0051	2.25×
Minmax-CF	27	−0.0010	2.55×

gistic regression and -0.0688 (-6.88 points) for HVG-500 MLP. In the latter setting, Random has higher mean ΔBA ($+0.0018$) and rare recall (0.984), while PCA is closer to Full (-0.0002). Across the 27 fold-seed results, Minmax-CF approximately retains average Full accuracy and reduces training time, although some held-out technologies show larger losses. The original IDs allow these folds to be checked for source-technology coverage; cell-level attribution requires the retained-ID outputs.

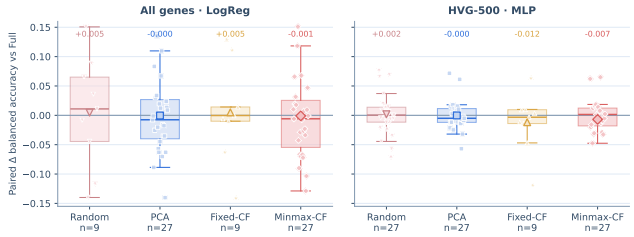

Figure 7: hPancreas fold-seed distributions of paired ΔBA relative to matched Full; horizontal zero denotes no difference. Technology-cluster bootstrap intervals are reported in the text, not drawn here, and cross zero for Minmax-CF in both downstream configurations.

Norman perturbation-support analysis. Overall group-delta Pearson correlation for Full/Random/PCA/Fixed-CF/Minmax-CF is $0.7572/0.3897/0.3385/0.3455/0.3597$. Minmax-CF improves over Fixed-CF but remains below Random; its pathway-delta MAE of 0.03465 is the best compressed value and is close to Full (0.03317). Stratifying by component visibility separates the support regimes. Minmax-CF reaches $r = 0.833$ when both perturbation components are observed in training (seen-2) and 0.537 for seen-1, but falls to -0.032 for seen-0 combinations and -0.025 for unseen singles. The other compressed selectors show the same qualitative transition.

Figure 4 shows this support transition. The official split determines component visibility: the known-component fraction is identical across methods (0.3897), and every compressed selector retains each nonempty training stratum. Full decreases from $r = 0.898$ for seen-2 combinations to $r = 0.700$ for seen-1 combinations. For seen-0 combinations and unseen singles, its correlation is undefined because the additive predictor has no train-derived estimate for the missing components. All compressed selectors show the same support-dependent drop, which also appears in Full and reflects limits of the training support and additive predictor. Relative to Full, Minmax-CF loses 0.064 correlation in seen-2 and 0.163 in seen-1. Because the retained examples are original cell IDs, this residual gap can be audited for underrepresentation of highly responsive cells, rare transcriptional states, batches, or guide-specific subpopulations within an observed perturbation. An ID-level audit is needed before residual error can be assigned to individual cells.

Table 3: IFNB joint cell–gene seed-0 pilot (MLP). BA and seconds are cell-type/stimulation; MB is peak memory. NCFM denotes the fixed-CF real-cell pilot selector.

Configuration	Cells × genes	BA ↑	Seconds ↓	MB ↓
Full+All	9799 × 14053	.845/.983	4.962/4.833	119.5
Full+Bio500	9799 × 500	.874/.990	1.262/1.262	66.0
NCFM50+All	1224 × 14053	.814/.951	.527/.527	119.5
NCFM50+Bio500	1224 × 500	.868/.963	.166/.163	66.0
PCA50+Bio500	1224 × 500	**.875/.972**	.163/.162	66.0
Random50+Bio500	1224 × 500	.866/.959	**.163/.161**	66.0

IFNB seed-0 cell–gene compression pilot. Table 3 isolates the second, gene-budget axis on IFNB. NCFM-50+Bio-500 reduces the training object from 9,799 to 1,224 cells and from 14,053 to 500 genes. Its MLP BA is 0.868 for cell type versus 0.845 for Full+All, while stimulation BA is 0.963 versus 0.983. Per-task downstream time is about 3.3% of Full+All and peak allocated memory is 44.8% lower. A masked-autoencoder proxy takes about 3.8% of Full+All pretraining time and 66.4% less peak memory. However, PCA-50+Bio-500 has higher pure classification BA ($0.875/0.972$ for cell type/stimulation) than NCFM-50+Bio-500. Because these results use only seed 0, we report IFNB separately from the cross-dataset OOD analysis.

Fidelity–Utility–Cost Analysis

Failure modes and auditability. Table 4 lists four settings where lower CF discrepancy is insufficient: unseen perturbation components, rare-state recognition, technology-specific shifts, and weak fidelity–utility coupling. On hPancreas, fold-seed soft-CF reductions correlate only weakly with ΔBA (Pearson $r = 0.29$ for all-gene logistic regression and -0.17 for HVG-500 MLP; Figure 5). The retained IDs allow each case to be checked against its training support, including perturbation-component visibility, within-component state coverage, rare-cell and donor retention, and source-technology composition. Cell-level error attribution and cluster-adjusted inference remain future work.

Cell-level audit design. The current Norman artifacts contain subgroup averages but omit the selected cell IDs, so cell-

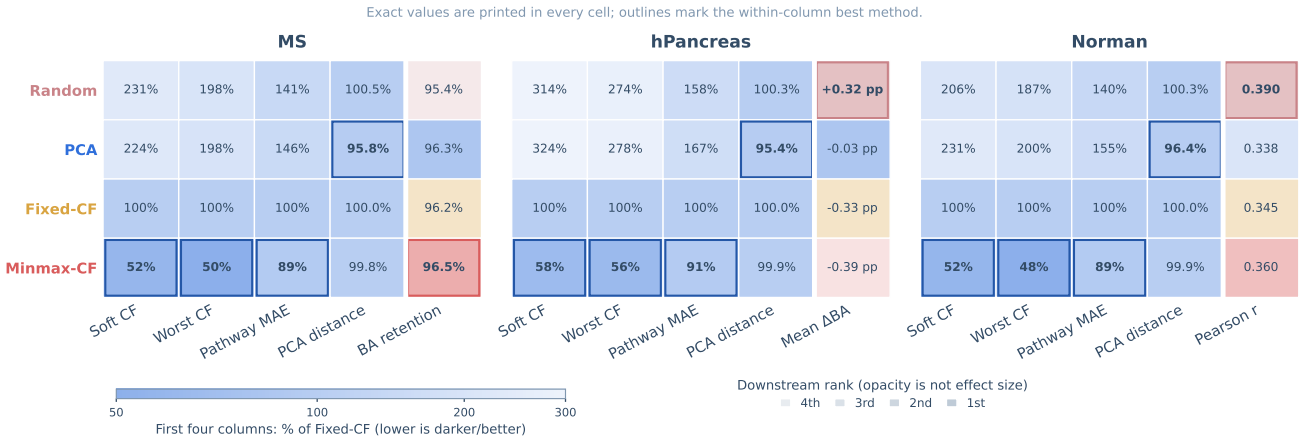

Figure 8: Preservation profiles of the four compressed selectors. The first four columns report each error as a percentage of Fixed-CF; darker blue means lower error, and exact values are printed in the cells. Downstream entries are mean matched-Full BA retention on MS, the equally weighted mean ΔBA across the two hPancreas configurations in percentage points, and group-delta Pearson correlation on Norman.

level attribution is not yet possible. An audit should export the selected IDs for every method and seed and join them to the original expression matrix and provenance metadata. Within each observed perturbation group, it should compare response strength, transcriptional-state coverage, batch and guide composition, quality-control fields, and retention of rare or highly responsive cells between the selected subset and Full. Each test combination can then receive a weakest-component coverage score, which can be related to condition-level Pearson correlation and pathway MAE within the seen-2 and seen-1 regimes. This association would be exploratory. A stronger test is a training-only rescue experiment that replaces selected cells with coverage-improving real cells at the same budget, without using held-out test labels. Reproducible recovery across seeds would link within-component underrepresentation more directly to prediction error.

Comparison with simple baselines. Random selection remains competitive at these budgets because it uses explicit biological stratification. Figure 8 reports the preservation trade-off without pooling incompatible units into one score. Minmax-CF ranks first on soft-CF, worst-CF, and pathway-mean error in all three datasets, whereas PCA ranks first on nearest-PCA distance. Downstream results differ by dataset: Minmax-CF narrowly leads MS retention, while Random leads hPancreas and Norman. The values printed in each cell also show the size of these differences. For example, Minmax-CF and PCA retain 96.5% and 96.3% MS BA, so their different rank colors correspond to a small absolute gap. Worst-direction matching and local manifold coverage capture different properties, and neither alone predicts responses to unseen causal components. A preregistered hybrid constraint could combine the two objectives using training and validation data only.

Selection and repeated-training cost. Minmax-CF selection takes 10.60 s on MS, 18.58 s per hPancreas fold on average, and 43.62 s on Norman, compared with 3.15, 4.49, and 4.45 s for Fixed-CF. Selection is paid once and can be amortized across downstream models, seeds, and tuning runs. For a single short classifier run, however, selector overhead may dominate. In all-gene logistic regression, median speedups range from $2.25\times$ to $2.84\times$ and mean paired ΔBA remains within 0.51 percentage points of Full. Figure 9 isolates repeated GPU training in the joint cell–gene HVG-500 MLP setting: Random has the highest mean utility, while Minmax-CF is the fastest compressed point at $2.09\times$. The connected line contains empirically non-dominated method means, and the wide fold–seed variability bars show uncertainty in the utility estimates. We report selection and training costs separately.

Regime	Observed signal	What retrieval makes inspectable
Perturbation support	Full r: $.898/.700$ for seen-2/1; Minmax-CF: $.833/.537/-.032/-.025$	Was a component absent under the official split, or were responsive within-condition states of an observed component underrepresented among retained cells?
Rare states	MS BA retention 96.52%; rare recall $0.704 < 0.90$ target	Which rare cells, donors, and source records were retained?
Technology shift	hPan ΔBA means $-.10/-.69$ pp; worst folds $-8.87/-6.88$ pp (LogReg/MLP)	Are retained states dominated by particular source technologies?
Objective mismatch	Soft-CF gain vs. ΔBA: Pearson $.29/-.17$ (LogReg/MLP)	Do CF-repair cells support the downstream decision boundary?

Table 4: Observed failure regimes and the diagnostic questions enabled by exact real-cell retrieval. Fold extremes and correlations are descriptive, not inferential bounds; retrieval makes data support inspectable but does not prevent failure or establish its cause.

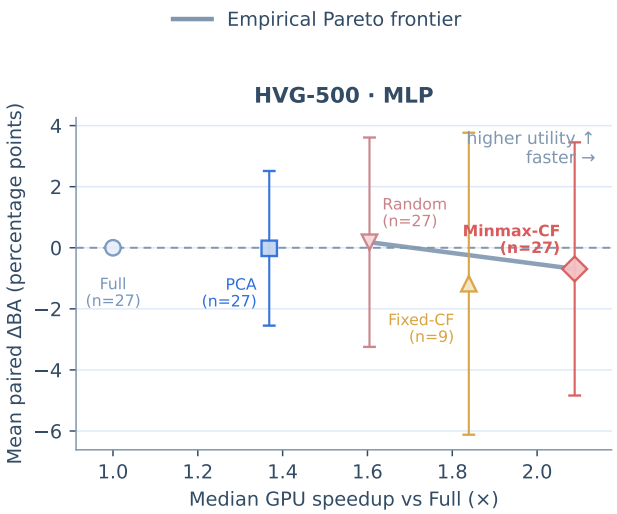

Figure 9: hPancreas HVG-500 MLP utility–training-efficiency trade-off. Points are method-level means at median GPU speedup relative to Full; vertical bars are standard deviations, not confidence intervals, of paired ΔBA across available fold–seed units. Labels report n. The line connects empirically non-dominated points and does not represent a training trajectory. One-time selector cost is excluded.

Limitations. First, the evaluation compares real-cell selectors under compact downstream models. It does not include a controlled head-to-head comparison with large single-cell foundation models. Second, Random and Fixed-CF have only one downstream seed in some hPancreas configurations, whereas PCA and Minmax-CF have three; Table 2 reports these unequal sample sizes.

Conclusion

We presented traceable single-cell data distillation in which every retained example is an original cell ID paired with original gene symbols. **Minmax-CF** combines this discrete constraint with worst-direction characteristic-function matching and reduces the target discrepancies on all three datasets. Across three out-of-distribution (OOD) regimes, the resulting coresets are competitive and can reduce repeated training cost. Worst-case-aware real-cell coresets provide an auditable alternative to synthetic prototypes. Their value lies in balancing fidelity, downstream utility, and computational cost, not in replacing full datasets or large pretrained models in every setting.

References

Ahlmann-Eltze, C.; Huber, W.; and Anders, S. 2025. Deep-Learning-Based Gene Perturbation Effect Prediction Does Not Yet Outperform Simple Linear Baselines. *Nature Methods*.

Cazenavette, G.; Wang, T.; Torralba, A.; Efros, A. A.; and Zhu, J.-Y. 2022. Dataset Distillation by Matching Training Trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10718–10727.

Gene Ontology Consortium. 2023. The Gene Ontology Knowledgebase in 2023. *Genetics*, 224(1): iyad031.

Hie, B.; Cho, H.; DeMeo, B.; Bryson, B.; and Berger, B. 2019. Geometric Sketching Compactly Summarizes the Single-Cell Transcriptomic Landscape. *Cell Systems*, 8(6): 483–493.e7.

McInnes, L.; Healy, J.; and Melville, J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*.

Milacic, M.; et al. 2024. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Research*, 52(D1): D672–D678.

Norman, T. M.; et al. 2019. Exploring Genetic Interaction Manifolds Constructed from Rich Single-Cell Phenotypes. *Science*, 365(6455): 786–793.

Persad, S.; et al. 2023. SEACells Infers Transcriptional and Epigenomic Cellular States from Single-Cell Genomics Data. *Nature Biotechnology*, 41: 1746–1757.

Roohani, Y.; Huang, K.; and Leskovec, J. 2023. Predicting Transcriptional Outcomes of Novel Multigene Perturbations with GEARS. *Nature Biotechnology*.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605.

Wang, S.; Yang, Y.; Liu, Z.; Sun, C.; Hu, X.; He, C.; and Zhang, L. 2025. Dataset Distillation with Neural Characteristic Function: A Minmax Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25570–25580.

Wang, T.; Zhu, J.-Y.; Torralba, A.; and Efros, A. A. 2018. Dataset Distillation. *arXiv preprint arXiv:1811.10959*.

Yu, Z.; Han, J.; Liu, Y.; and Chen, Q. 2025. scDD: scRNA-seq Dataset Distillation in Latent Codes with Single-Step Conditional Diffusion Generator. *Knowledge-Based Systems*, 330: 114610.

Zhao, B.; and Bilen, H. 2023. Dataset Condensation with Distribution Matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6514–6523.

Zhao, B.; Mopuri, K. R.; and Bilen, H. 2021. Dataset Condensation with Gradient Matching. In *International Conference on Learning Representations*.