Sample-Efficient Deep Cellular Deconvolution with Bulk-Anchored Pseudo-Bulk Augmentation

Anonymous authors
Paper under double-blind review

Abstract

Cellular deconvolution estimates cell-type proportions from bulk RNA-seq and enables cell-population analysis across large transcriptomic cohorts. Most methods depend on a cell-type-resolved reference, typically from scRNA-seq; when that reference is external to the target bulk cohort or differs in processing, protocol, or disease context, the mismatch reduces accuracy, especially for closely related cell states. Paired bulk and scRNA-seq data avoid this by directly connecting the target bulk assay to matched single-cell-derived composition, but large-scale paired profiling is impractical, so in realistic studies paired data are available only for a limited pilot subset of donors. This creates a challenge for deep deconvolution, which learns flexible bulk-to-composition mappings but needs enough labeled examples: the usual fix, training on pseudo-bulks generated from the reference, does not exploit the real paired bulk profiles a pilot design provides. We address this with an anchored data-augmentation strategy: from the matched in-study scRNA-seq we generate pseudo-bulks to expand the number and diversity of labeled examples, and we retain the real paired bulks in training with their matched composition labels, anchoring the model to the target bulk RNA-seq distribution. This converts a small paired pilot into an expanded supervised training resource. On a paired bronchoalveolar-lavage benchmark, the strategy makes a standard deep deconvolver the best method at every cell-type resolution, ahead of the strongest classical method and standard pseudo-bulk training, and controlled experiments show the real-bulk anchors drive the gain. Code is available at `https://anonymous.4open.science/r/2dRNA-512F`.

1 Introduction

A tissue sample is a mixture of many different cell types, and much of its biology is governed not by which genes are active in isolation but by which cells are present and in what proportions. An influx of immune cells, the loss of a structural population, or a shift in the balance between closely related subtypes are all compositional changes that track how a tissue responds to disease. The workhorse assay for reading a sample's molecular state, bulk RNA-seq, measures the *aggregate* gene expression of the whole sample as a single averaged signal, summing over every cell at once. It is inexpensive and ubiquitous, but it discards the very information we often care about. The same bulk profile can be produced by quite different underlying cell mixtures.

Cellular deconvolution is the inverse problem of recovering the cell-type proportions from that aggregate signal. Solving it lets researchers reuse the large amounts of bulk RNA-seq already collected to ask how cell-type composition shifts between healthy and diseased samples, without the expense of profiling every sample at single-cell resolution. Two families of methods dominate. Linear methods (CIBERSORT (Newman et al., 2015), MuSiC (Wang et al., 2019), EPIC (Racle et al., 2017), CIBERSORTx (Newman et al., 2019)) express a bulk profile as a weighted sum of fixed per-cell-type reference signatures. Deep-learning methods, the most prominent being Scaden (Menden et al., 2020), instead train a neural network to map a bulk expression vector directly to a vector of cell-type fractions.

Training such a network requires labeled examples: expression profiles paired with known cell-type fractions. Real bulk samples rarely come with such labels, so deep deconvolvers commonly manufacture training data from a single-cell RNA-seq reference. Because single-cell sequencing assigns a cell-type label to every individual cell, one can select a collection of cells, sum their expression vectors into a synthetic "pseudo-bulk" sample, and use the known composition of that collection as the training label. Repeating this thousands of times yields a large labeled corpus on which the network learns the expression-to-composition mapping.

This pseudo-bulk strategy is useful, but its quality depends on the reference from which the pseudo-bulks are generated. In many applications, the available single-cell reference is external to the target bulk cohort or differs in sample handling, sequencing protocol, disease context, or cell-state composition. Such mismatch can distort the relationship between reference-derived expression profiles and the real bulk RNA-seq samples to which the model will be applied. Recent paired bulk + single-cell benchmarking in human bronchoalveolar lavage (BAL) showed that even a small protocol-matched paired subset can improve deconvolution by providing an in-study reference and by enabling paired-data-based filtering of genes that behave discordantly between bulk and single-cell-derived profiles (Hu et al., 2026). These results show that paired data contain information about the target study that is not available from unmatched references alone. So far that information has been put to work only to choose a better reference and to filter discordant genes; we ask whether it can do more, and directly improve how a deep deconvolution model is trained.

The practical limitation is that paired profiling is difficult to obtain at scale. Bulk RNA-seq can be generated for large cohorts, but single-cell RNA-seq is more expensive, more technically demanding, and more sensitive to sample handling. Thus, in realistic studies, paired bulk + single-cell data are more likely to be available for a limited pilot subset than for the entire cohort. This creates a specific challenge for deep deconvolution: paired samples provide high-quality supervised examples, because each real bulk profile can be linked to a matched single-cell-derived composition label, but the number of such examples is usually too small to train a flexible neural network directly.[1]

We address this limited-pair setting by constructing a larger supervised training set from the paired pilot cohort. The strategy uses two sources of training information. First, pseudo-bulk samples generated from the matched in-study single-cell data expand the number and diversity of labeled examples available to the neural network. Second, the real paired bulk samples are retained in the same training set with their matched single-cell-derived composition labels. These real bulk examples anchor training to the target bulk RNA-seq distribution, while the pseudo-bulk examples provide augmentation. In implementation, real bulk and pseudo-bulk inputs are harmonized to the shared gene space so that they can be used jointly during supervised training.

Hu & Chikina (2024) showed that donor preservation matters when pseudo-bulks are used for evaluating deconvolution methods: methods can rank differently under donor-preserved versus donor-pooled simulated bulks. Here we study a related but distinct question in the paired training setting: when only a limited number of paired donors are available, can pseudo-bulk augmentation from the matched single-cell data, combined with real paired bulk examples, improve deep deconvolution on real bulk RNA-seq? We find that it can.

Contributions. Our central finding is that, in the paired setting, careful construction of the training data can substantially improve deep deconvolution. The proposed strategy converts a small paired bulk + single-cell pilot cohort into an expanded supervised training resource: pseudo-bulks increase effective sample size, while real paired bulks keep the model calibrated to the target bulk assay. The trained model requires only a bulk RNA-seq profile at inference time, so the paired single-cell data are needed only for training. We support this with controlled experiments that vary how the training data are constructed while holding the network fixed. Concretely:

[1] We use *sample* in the biological sense throughout, meaning one donor measured by both assays: the scarce and costly unit our strategy economizes, and the sense intended by "sample-efficient" in the title. This is distinct from the machine-learning usage, where a sample is a single training example; in that sense our augmented training set is deliberately large, assembled from a small number of paired donors.

- **An anchored data-augmentation strategy for limited paired data** (§4). From the paired single-cell data, we generate pseudo-bulk samples to augment the small number of donors, and we keep the real paired bulk samples in the training set with their matched single-cell-derived composition labels. This applies to any network that maps a bulk expression vector to cell-type fractions.
- **Evidence that the real-bulk anchor drives the gain.** Controlled experiments that hold the network fixed and vary only the training data show the anchor is load-bearing: against a fair composition-broad augmented baseline, keeping the real bulks lowers mean MSE by 35 to 94% across architectures, and a pseudo-only ablation that removes them worsens MSE by 1.4 to 5.8× (§6.2). The augmentation, for its part, helps mainly by exposing the network to a broad range of compositions. Further, drawing each pseudo-bulk within a single donor (donor-stable) adds a secondary gain that is real and significant for the direct-mapping MLPs (§6.3).
- **Improved performance on a real paired BAL benchmark.** Applied to the Scaden architecture, the paired-training strategy attains the best mean MSE, RMSE, correlation, and R^2 at every granularity level on the paired BAL benchmark, against sixteen published method variants. It significantly beats the strongest classical method (CSCD) and standard pseudo-bulk deep training, and even our deliberately small network is on par with CSCD (§6.1, §6.4).

The effect is not universal. How much it helps depends on the architecture, and §6.3 maps where it fades.

Paired cohorts, with bulk and single-cell measured on the same donors, are scarce. We work in this limited but increasingly important regime, where the central question is how to use a small number of paired samples most effectively for deep deconvolution. The small scale is intrinsic to the setting rather than a shortcoming of the study: paired profiling is expensive, so the fact that a pilot-scale cohort already yields strong deconvolution is the finding of interest, and an argument for collecting more paired data rather than a reason to discount this one. The evidence below comes from a single 30-donor BAL cohort, where we establish the claims *for this paired-data setting*. Whether the same training-data construction strategy transfers to other tissues and larger cohorts is left to future work, as more paired data become available (§7).

2 Related Work

Classical deconvolution. The dominant classical approach writes a bulk profile as a weighted sum of fixed per-cell-type reference signatures. CIBERSORT (Newman et al., 2015) introduced support-vector regression for immune-cell fractions; CIBERSORTx (Newman et al., 2019) added batch correction and signature construction from single-cell references; BSEQ-sc (Baron et al., 2016) built such signatures directly from single-cell data; and MuSiC (Wang et al., 2019) introduced a multi-subject reference to handle cross-donor heterogeneity. Later methods refine marker selection (dtangle (Hunt et al., 2019), EPIC (Racle et al., 2017), TIMER2 (Li et al., 2020), and the multi-objective gene selection of AutoGeneS (Aliee & Theis, 2021)) or reweight genes to remove cell-abundance bias (DWLS (Tsoucas et al., 2019)). A parallel reference-free line instead infers signatures and proportions jointly (CDSeq (Kang et al., 2019)). A recent benchmark (Avila Cobos et al., 2020) compares these methods head-to-head. The classical baselines in our evaluation are drawn from this family (§5). Because none of them trains a model on the synthetic pseudo-bulks we construct, there is no training-data-construction step for our strategy to change; it applies only to learned deconvolvers, and these classical methods are what ours are measured against.

Deep deconvolution. Scaden (Menden et al., 2020) showed that a neural network trained on a large corpus of synthetic pseudo-bulks can beat linear methods; it is the canonical deep deconvolver and our deep baseline. Most later deep methods, including recent ones such as DeSide (Xiong et al., 2024), refine the architecture or loss but inherit Scaden's recipe for the pseudo-bulks themselves, drawing cells by cell type without regard to donor identity and summing them, which on a multi-donor reference mixes cells across donors. Some scale that synthetic data drastically; Kassandra (Zaitsev et al., 2022), for instance, trains a gradient-boosted model on millions of artificial transcriptomes, yet still builds those mixtures without preserving donor identity. Our strategy departs from this recipe on two axes, keeping the real paired bulks in training as anchors and controlling the composition diversity of the pseudo-bulks. Both concern how the

training data is built, not how a method corrects batch or cross-dataset effects at test time, a separate and complementary axis.

Bridging the synthetic-to-real gap by adapting the model. Many methods keep the cross-donor training recipe and instead correct the resulting train/test mismatch downstream. Some harmonize or batch-correct the expression inputs. Others adapt the trained model to the target domain at inference. TAPE uses a tissue-adaptive autoencoder (Chen et al., 2022), and DISSECT adds a semi-supervised consistency loss that ties simulated and real samples together (Khatri et al., 2024). A third group learns domain-invariant or generative representations, as in BuDDI's variational autoencoder (Davidson et al., 2025) and MixupVI's mixup-regularized latent (Grouard et al., 2025). Context-specific deep deconvolvers such as digitalDLSorteR (Torroja & Sanchez-Cabo, 2019) tailor a model to each dataset. All of these adapt the *model*; we change the *data* the model is trained on. The two axes are orthogonal and could be combined.

Donor preservation. Hu & Chikina (2024) showed that pooling cells across donors produces pseudo-bulks with artificially low biological variance, whose cell-type expression profiles collapse toward the global mean, and that methods rank differently when *evaluated* on donor-preserved versus donor-pooled simulations. A recent systematic review (Lomas Redondo et al., 2025) catalogs deep deconvolvers by whether their pseudo-bulks come from a single individual or from many pooled together.[2] The distinction is recognized, but as an incidental property of a method's inputs, and no prior deep deconvolver makes donor preservation an *enforced* construction choice or *isolates* its effect with a controlled experiment. We do both, and find that once composition coverage is matched, donor preservation is a real but secondary effect rather than the main driver of the gap (§6.3); our central move is instead the use of paired real bulks as training anchors.

Which reference, versus how it is built. A separate line asks not how the pseudo-bulks are built but *which reference* to build them from. The benchmark we adopt (Hu et al., 2026) shows that a small, protocol-matched in-study reference beats much larger external ones, and SCDC (Dong et al., 2021) ensembles several single-cell references to average out batch effects. Both are about *which* reference to use. Our strategy acts one level deeper, on *how* the training data is built from a given reference, an axis they leave fixed. Closest to our main choice, keeping real bulks in training, is DAISM-DNN (Lin et al., 2022), which blends a few real calibration samples into its in-silico mixtures, anticipating our use of real bulks as anchors. DAISM uses the technique but does not isolate its effect; our controlled add/remove experiments (§6.2) show the real bulks are the load-bearing ingredient. DAISM's own behavior fits this account, since it already anchors, donor-stable sampling adds modestly to it (§6.3). Paired bulk + single-cell data is otherwise rarely used for training, simply because it is scarce.

Paired benchmark. Our testbed is the paired bronchoalveolar-lavage benchmark of Hu et al. (2026), with its fixed 10-fold split and panel of classical baselines. The two papers contribute on different axes. Theirs benchmarks existing classical methods and recommends which reference to use, proposing no deep-learning method; ours introduces a training strategy for deep deconvolvers and isolates why it works.

3 Problem setting

3.1 Setup and notation

We work in the *paired* bulk + single-cell setting, where each donor is measured twice on the same tissue, once in bulk and once at single-cell resolution. Throughout, the problem is sized by N donors, G genes, and k cell types. Write $\mathcal{D} = \{1, \ldots, N\}$ for the donors and $\mathcal{K} = \{1, \ldots, k\}$ for the cell types. Each donor $d \in \mathcal{D}$ is measured by two separate assays. Bulk RNA-seq gives a profile $b_d \in \mathbb{R}^G_{\geq 0}$ by sequencing the pooled mRNA of the whole tissue at once. Single-cell RNA-seq gives a sample of n_d cells, which we carry as two matrices, an expression matrix $E_d \in \mathbb{R}^{G \times n_d}$ with one column per cell, and a one-hot label matrix $Y_d \in \{0, 1\}^{k \times n_d}$

[2] The review marks a few methods, Scaden among them, as single-individual. That reflects how each method's original study happened to split its input files, not a rule the method enforces. Scaden's generator samples cells by cell type alone and never reads a donor label, so a pseudo-bulk is single- or cross-donor purely by accident of the input, and on the multi-donor references used in practice it pools across donors.

whose column for each cell marks its type in $\mathcal{K}$. Both assays use the same G genes, the set shared by the two sequencing platforms (§4).

The donor's *composition* is simply the fraction of its cells of each type. In matrix form, multiplying the one-hot label matrix Y_d by the all-ones vector $\mathbf{1}$ adds up its columns, one per cell, to give the number of cells of each type, and dividing by the cell count n_d turns those counts into fractions,

$$f_d \;=\; \frac{1}{n_d}\, Y_d\, \mathbf{1} \;\in\; \Delta^{k-1}, \qquad \text{i.e.} \quad f_d[j] \;=\; \frac{\#\{\text{donor } d\text{'s cells of type } j\}}{n_d}. \tag{1}$$

The result is a point on the probability simplex $\Delta^{k-1} = \{p \in \mathbb{R}^k_{\geq 0} : \sum_j p[j] = 1\}$, a length-$k$ vector of non-negative entries summing to one. *Deconvolution* is the task of learning a map $h_\theta : \mathbb{R}^G_{\geq 0} \to \Delta^{k-1}$ with $h_\theta(b_d) \approx f_d$. At inference only a bulk vector is needed; the single-cell side is used solely to build training data.

A linear mixing model underlies both the problem and classical deconvolution. From the single-cell side, let $s_{d,j} \in \mathbb{R}^G$ be the mean expression of donor d's type-j cells (the columns of E_d labelled j), its *signature* for type j, and let $S_d = [\, s_{d,1}, \ldots, s_{d,k} \,]$ stack these k signatures as columns. Under this mixing model, the donor's bulk is approximately a proportion-weighted combination of those signatures,

$$\underbrace{\begin{bmatrix} b_{d,1} \\ \vdots \\ b_{d,G} \end{bmatrix}}_{G\times 1} \approx \underbrace{\begin{bmatrix} | & | & & | \\ s_{d,1} & s_{d,2} & \cdots & s_{d,k} \\ | & | & & | \end{bmatrix}}_{G\times k \,=\, S_d} \underbrace{\begin{bmatrix} f_d[1] \\ \vdots \\ f_d[k] \end{bmatrix}}_{k\times 1 \,=\, f_d}, \tag{2}$$

that is, $b_d \approx S_d\, f_d$. This is a *model* of b_d, not a definition of it, since the bulk is produced by its own separate assay. Stacking all N donors column by column gives the cohort-level bulk and composition matrices

$$B = [\, b_1, \ldots, b_N \,] \in \mathbb{R}^{G\times N}, \qquad F = [\, f_1, \ldots, f_N \,] \in \mathbb{R}^{k\times N}.$$

The dominant classical approach assumes a single shared signature matrix explains the whole cohort, $B \approx SF$, and inverts it. Paired data shows this is too strong in two ways. (i) Each donor carries its own signatures S_d, so the cohort is N separate factorizations $b_d \approx S_d f_d$, not one shared one. (ii) Even those per-donor relations hold only approximately, since the bulk and single-cell assays measure the same tissue differently, a systematic offset rather than random noise. A classical inverter, tied to a single fixed S, can absorb neither. A learned deconvolver can, fitting a flexible mapping trained on the real bulks rather than inverting a fixed dictionary. Of the two, the systematic offset matters most for training and is what our real-bulk anchor targets (§3.2); the donor-to-donor variation in S_d also contributes, though only secondarily (§6.3).

3.2 Building the training set: pseudo-bulks and real anchors

Real bulks almost never come with their ground-truth fractions, so deep deconvolvers manufacture labelled training data as *pseudo-bulks* from the single-cell side. Concatenating the per-donor blocks of §3 gives the pooled cohort matrices $E = [\, E_1 \mid \cdots \mid E_N \,] \in \mathbb{R}^{G\times n}$ and $Y = [\, Y_1 \mid \cdots \mid Y_N \,] \in \{0,1\}^{k\times n}$ ($n = \sum_d n_d$). A pseudo-bulk selects a set C of cell columns; its input sums their expression and its label is their composition,

$$x_C \propto E_{:,C}\, \mathbf{1} \;\in\; \mathbb{R}^G_{\geq 0}, \qquad \pi_C = \frac{1}{|C|}\, Y_{:,C}\, \mathbf{1} \;\in\; \Delta^{k-1}. \tag{3}$$

A sampler q over C determines which cells enter each pseudo-bulk, and repeated draws give a labelled corpus $\mathcal{P} = \{(x_C, \pi_C)\}$.

A pseudo-bulk is a clean sum of single cells, so it carries none of the systematic offset that separates a real bulk b_d from its idealized mixture $S_d f_d$ of (2). A model trained on pseudo-bulks alone therefore never sees this offset and stays miscalibrated to the real bulks it will actually be applied to; consistent with this, benchmarking finds performance on pseudo-bulks a poor proxy for performance on real bulk (Hippen et al.,

2023). Paired data lets us fix this directly. Each real bulk comes with a matched single-cell-derived label f_d, so we keep the real pairs $\mathcal{A} = \{(b_d, f_d)\}$ in the training set as *anchors* beside the pseudo-bulks, teaching the model the offset instead of hiding it. Training minimizes the loss over both,

$$\min_{\theta} \sum_{d=1}^{N} \ell\big(h_\theta(b_d),\, f_d\big) \;+\; \sum_{(x_C,\pi_C)\in\mathcal{P}} \ell\big(h_\theta(x_C),\, \pi_C\big), \tag{4}$$

with ℓ the mean-squared error. The first sum, over the N real anchors, ties training to the real-bulk distribution; the second, over the pseudo-bulks, supplies the volume and composition diversity a handful of real pairs cannot. §6.2 shows the anchor term is what carries the benefit; the sampler q behind the augmentation term is what the rest of this section dissects.

Two properties of the sampler q matter for training. The first is the *composition distribution* it induces over the labels π_C, whether the pseudo-bulks cover a broad range of compositions or cluster in a narrow band. The second is whether the cells in a single C come from one donor or many. *Cross-donor* (q_{CD}), the recipe of prior deep deconvolvers, draws C from the pooled matrix E (training donors only, to prevent leakage), so one pseudo-bulk can mix cells from many donors; *donor-stable* (q_{DS}, ours) draws C from a single donor's E_d, so every pseudo-bulk stays within one donor. The two properties are entangled in the usual implementation, since sampling uniformly from the pooled cells also concentrates the compositions near the pool average; the controlled experiment of §6.3 separates them and finds the composition property carries most of the effect.

4 Method

Our *strategy* builds the paired-setting training set from real-bulk anchors and pseudo-bulk augmentation, with a gene-set restriction for compatibility, and works with any model that maps a bulk vector to cell-type fractions.

The training set is what we vary; the network is held fixed. Most of the paper uses one deliberately small multilayer perceptron (MLP), which we call `SimpleDNN`. It has two hidden layers (widths 128 and 64) with ReLU and dropout, a softmax output, Adam, and mean-squared-error loss, and we train it with early stopping on a held-out split inside each fold. We keep it small on purpose. Larger networks overfit the roughly 25-donor inner training set (the 27 training donors per fold, less the early-stopping split), so a small one makes the point cleanly. Any gain comes from the data, not from added capacity. We then repeat the experiment on two published networks (§6.3), Scaden (Menden et al., 2020) and DAISM-DNN (Lin et al., 2022), showing the effect is not specific to our own network.[3]

Of the three ingredients below, two are design choices we test empirically (§6.3, §6.2), both motivated by the structure of paired data; the third, the gene-set restriction, is a technical requirement rather than a choice.

Real bulks are kept as anchors. The real paired bulks stay in the training set, each with its matched single-cell-derived composition label, added *alongside* the pseudo-bulks rather than in place of them. These anchors are the load-bearing ingredient (§6.2). Removing them and training on pseudo-bulks alone degrades MSE by 1.4 to 5.8× across the architectures tested.

Pseudo-bulk augmentation. Alongside the anchors we add pseudo-bulks to expand the number and composition diversity of the labelled examples. For each training donor we generate 30 pseudo-bulks; to build one we redraw a random fraction $\alpha \sim U(0.3, 1.0)$ of *that donor's* cells and sum them into x_C as in §3.2. Drawing each pseudo-bulk from a single donor (donor-stable) is our default; the alternative used by prior deep methods, such as Scaden (Menden et al., 2020) as published, pools cells across donors. This choice matters less than the anchor or the composition coverage (§6.3), so we adopt the donor-stable default without leaning on it.

[3]We reimplement these networks ourselves, matching each one's architecture, loss, hyperparameters, optimizer, and input normalization from the authors' released code, so that every network trains on the identical pseudo-bulk pools within a single pipeline. The released implementations do not support swapping in our pools, so running the original code is not an option.

Gene-set restriction. Bulk RNA-seq measures roughly 62,000 genes; single-cell RNA-seq measures roughly 14,000. We restrict the model's input to the ∼14,000 genes that appear in both. The restriction is required because the pseudo-bulks have no expression for the bulk-only genes, so the two cannot otherwise be combined in the same training batch. It also cuts input dimensionality by ∼77%, which helps the small network avoid overfitting to uninformative features.

The remaining implementation details, including the sparse-matrix arithmetic used for efficient generation, are in the released code.[4]

5 Experimental Setup

Dataset. The paired BAL benchmark of Hu et al. (2026) provides 30 paired bulk + single-cell donors from bronchoalveolar lavage (BAL), 4 COPD-positive. The bulk arm covers 61,578 genes (TPM-normalized); the single-cell arm 14,378 genes across 241,924 cells (sparse CPM). Three cell-type granularity levels give $k = 11/17/13$ cell types at `celltype1` / `celltype2` / `celltype3` (abbreviated `ct1` / `ct2` / `ct3`) respectively. The `celltype2` level splits Macrophages and Monocytes into finer subtypes; one of its 17 classes (`Macrophage_MT1G`) is empty in this cohort, leaving 16 populated (see Table 1).

Cross-validation. The 10-fold leave-3-out split of Hu et al. (2026) assigns 30 donors deterministically to 10 folds of 3. Per fold, 3 donors are held out, 27 are used for training. We use the same split for all our configurations and all baselines.

Baselines. We compare against sixteen baseline method × marker-list combinations, pre-evaluated on the same CV split by Hu et al. (2026). They span ten methods, CSCD (Karimnezhad, 2022), CSsingle (Shen et al., 2026), Bisque (Jew et al., 2020), MuSiC (Wang et al., 2019), BayesPrism (Chu et al., 2022), EPIC (Racle et al., 2017), PREDE (Qin et al., 2020), fardeep (Hao et al., 2019), DeconRNASeq (Gong & Szustakowski, 2013), and TOAST (Li & Wu, 2019). Six of them (CSCD, EPIC, MuSiC, PREDE, Bisque, fardeep) are run with both a "robust marker" (`.F`) variant and the method default; the remaining four (CSsingle, DeconRNASeq, BayesPrism, TOAST) with the default only. We restrict the comparison to the `sc.ubchli` (paired single-cell) reference. Our method is trained on paired data, so the like-for-like comparison is to baselines using the same reference.

Metrics. Per held-out sample, we compute four metrics across the k predicted vs. ground-truth fractions, mean squared error (MSE), root MSE (RMSE), the coefficient of determination ($R^2 = 1 - \mathrm{SS}_{\mathrm{res}}/\mathrm{SS}_{\mathrm{tot}}$), and Pearson correlation (Cor). Per granularity level, we report the mean over the 30 outer-test samples. We recompute every metric ourselves, the baselines' included, from each method's saved predictions. One caveat applies to R^2. The source benchmark reports squared Pearson correlation under that name, so we use the coefficient of determination throughout for a like-for-like comparison. Our ranking is the same under squared correlation.

Significance. One-sided paired Wilcoxon signed-rank tests on per-sample metrics. Wilcoxon is preferred over the paired t-test because per-sample MSE distributions are right-skewed. A few outlier donors (typically the COPD cases) dominate the variance.

Configurations and seeds. Every result uses the single canonical configuration of §4 (donor-stable pseudo-bulks + real bulks + sc-overlap gene filter, MSE loss, no mixup, no KL). The headline benchmark (Table 1) averages three seeds $\{42, 7, 17\}$; analyses that need a fixed prediction set (the significance tests and the consistent-metric leaderboard) use the seed-42 run, which tracks the three-seed mean to within the 5 to 8% multi-seed coefficient of variation. Fold-to-fold variability is far larger (60 to 96% CoV across the 10 folds), because a single COPD outlier among a fold's 3 held-out donors can dominate it. This is why our

[4] We also tried two further design choices, mixup (Zhang et al., 2018) over training batches (linearly interpolating pairs of training rows, a simpler use than baking mixup into a generative latent as in MixupVI (Grouard et al., 2025)), and KL-divergence loss in place of MSE. Neither added measurable improvement over the components above, so we omit both from the method; Appendix B dissects the pseudo-bulk-pool × regularization interaction that this rests on.

central comparisons (the A/B of Table 4 and the significance tests) are *paired* on the same held-out donors, so the large cross-fold variance cancels instead of swamping the signal.

6 Results

This paper's contribution is a training strategy, not a new architecture, and we establish it in three steps. We lead with the benchmark (§6.1) to show the practical payoff. A stock Scaden trained with our strategy is the best method at every granularity level, and even `SimpleDNN`, our deliberately small network, is on par with the strongest classical baseline. We then isolate why it works. A controlled experiment (§6.2) shows the real bulks are the load-bearing ingredient. Finally we dissect the augmentation (§6.3). An ablation confirms it drives most of the non-anchor gain, and a controlled experiment traces that gain to composition diversity. Throughout, architecture, gene set, schedule, and seeds are held fixed and only the training data varies, so any gain is a property of how the training data is built.

6.1 Our strategy sets the best benchmark result at every level

Trained with our strategy, Scaden is the best method on the benchmark. It has the lowest mean MSE and RMSE and the highest Cor and R^2 at every granularity level (Table 1), ahead of all sixteen classical variants and of standard pseudo-bulk deep training. Its multi-seed coefficient of variation on MSE is 5 to 8%, so the ranking does not depend on the random seed, and its margin over CSCD, the strongest classical method, is significant at every level (paired Wilcoxon $p \leq 0.012$; §6.4). The full per-sample distributions behind these means, for all eighteen methods on every metric and level, are in Appendix A (Figure 1).

Table 1: Three deep architectures trained with our strategy (marked "+ ours") vs. the strongest classical method (CSCD) at each granularity level on the `sc.ubchli` reference, sorted by mean MSE. Deep rows are three-seed means; the best value per column is bold. The `ct2` average includes one all-zero class (`Macrophage_MT1G`) that scales every method's `ct2` error equally and leaves rankings unchanged. The `ct2` sample-count caveat, seed stability, and the R^2 convention are discussed in the text and §5.

Level	Method	mean Cor	mean MSE	mean R^2	mean RMSE
`ct1` (k=11)	**Scaden + ours**	**0.9831**	**0.00197**	**0.9463**	**0.0360**
	DAISM-DNN + ours	0.9802	0.00255	0.9312	0.0398
	SimpleDNN + ours	0.9809	0.00295	0.9197	0.0430
	CSCD (best classical)	0.9784	0.00327	0.9104	0.0454
`ct2` (k=17)	**Scaden + ours**	**0.9645**	**0.00151**	**0.9018**	**0.0328**
	DAISM-DNN + ours	0.9548	0.00199	0.8732	0.0385
	SimpleDNN + ours	0.9522	0.00224	0.8580	0.0409
	CSCD (best classical)	0.9408	0.00253	0.8054	0.0440
`ct3` (k=13)	**Scaden + ours**	**0.9765**	**0.00202**	**0.9151**	**0.0354**
	DAISM-DNN + ours	0.9696	0.00261	0.8923	0.0433
	SimpleDNN + ours	0.9702	0.00305	0.8721	0.0456
	CSCD (best classical)	0.9660	0.00349	0.8489	0.0507

The `ct2` comparison carries one caveat. At `ct2`, CSCD returns valid predictions for only 24 of the 30 samples, so its `ct2` numbers and the paired `ct2` test cover that subset only. The dropped samples are exactly the ones for which CSCD returned no valid prediction, so excluding them can only help CSCD; our `ct2` comparison is, if anything, conservative.

Our strategy lifts the other two networks as well. DAISM-DNN is the second-best deep configuration and `SimpleDNN` the third, and both beat every classical method on mean MSE. Even `SimpleDNN`, the least accurate of our three feed-forward nets, is competitive with CSCD. Its mean MSE is roughly 10 to 13% below CSCD's. The per-sample paired test calls that a tie at all three levels ($p > 0.05$; §6.4, Table 6),

because the gap is small (about 4×10^{-4}) and a few COPD outliers dominate the variance. Against the other fifteen classical variants the same network wins outright.

Taken together, these results point to the training data, not a new architecture, as what moves the result on this paired BAL benchmark. The best configuration pairs an off-the-shelf network (Scaden) with our strategy, and even a small network is on par with the strongest classical method. This benchmark shows the strategy works in practice; the analysis that follows isolates why.

6.2 Real bulks as anchors

The real-bulk anchors are what drive the strategy, and two controls establish this from opposite directions, adding them to a fair augmented baseline and removing them from the full pipeline.

Adding anchors to a fair baseline. The cleanest test pits the anchor against a competent augmentation. We build a *fair* augmented baseline whose pseudo-bulks already span a broad composition range, drawn cross-donor from the pooled cells with broad random compositions (half a complete Dirichlet, half sparse), matched to our augmentation in count (810/fold) and per-sample size but carrying no donor structure. To this base we add the real bulks as training anchors, and then switch its pseudo-bulks to donor-stable to recover our full configuration. Table 2 reports all three.

Table 2: Adding real-bulk anchors to a fair, composition-broad augmented baseline drives the gain (three-seed mean MSE). Row (1) is the fair baseline with no anchors; (2) adds the real bulks; (3) additionally makes the pseudo-bulks donor-stable (our full configuration). Anchoring, (1)→(2), is the dominant step; donor-stable, (2)→(3), adds a small residual. DAISM anchors on real bulks by construction, so it has no no-anchor row.

Architecture	Configuration	`ct1`	`ct2`	`ct3`
SimpleDNN (ours)	(1) fair, no anchor	0.00690	0.04880	0.00991
	(2) fair + anchor	0.00448	0.00280	0.00348
	(3) full (donor-stable + anchor)	**0.00295**	**0.00224**	**0.00305**
Scaden	(1) fair, no anchor	0.00633	0.02121	0.00804
	(2) fair + anchor	0.00213	0.00185	0.00276
	(3) full (donor-stable + anchor)	**0.00189**	**0.00157**	**0.00207**
DAISM-DNN	(2) fair + anchor	0.00291	0.00230	0.00275
	(3) full (donor-stable + anchor)	**0.00255**	**0.00199**	**0.00261**

Adding the real bulks to the fair baseline, (1)→(2), lowers MSE by 35 to 94% for `SimpleDNN` and 66 to 91% for Scaden (the top of each range is `ct2`, where the unanchored model fails especially hard). The anchor is the dominant ingredient, and the effect holds against a competent broad-composition baseline, not only the narrow pool-average one. Switching the pseudo-bulks to donor-stable, (2)→(3), then adds a further ∼12 to 34% (`SimpleDNN`), ∼11 to 25% (Scaden), and ∼5 to 13% (DAISM-DNN), a small residual that the donor-coupling analysis of §6.3 isolates independently.

Removing anchors from the full pipeline. The complementary control drops the 27 real bulks per fold from the training set, leaving only the 810 pseudo-bulks (donor-stable or cross-donor) per fold.

Removing the real bulks moves training back into the degraded regime (Table 3). The cross-donor pseudo-only rows use no real bulks and pool cells across donors, the closest analog to standard Scaden-style training data; at every level they are 1.8 to 5.8× worse than the full pipeline for `SimpleDNN` and 1.4 to 1.6× worse for Scaden. For `SimpleDNN` the donor-stable pseudo-only row isolates the loss to the anchor itself, dropping the real bulks but keeping the composition-broad donor-stable pool, so the 2.3× degradation is not an artifact of cross-donor pooling. Both controls point the same way. On this paired benchmark the augmented training set depends on the real bulks; without them it collapses back toward standard pseudo-bulk training.

Table 3: Removing the real bulks from training degrades mean MSE for both architectures (three-seed means). The "Full" rows keep both real bulks and donor-stable pseudo-bulks; the pseudo-only rows drop the real bulks. The final column is the average degradation, the mean factor by which a pseudo-only row's per-level MSE exceeds its block's Full row; the Full value per block is bold. DAISM-DNN is omitted because it anchors on real bulks by construction and so has no anchor-free (pseudo-only) form. `SimpleDNN` is shown with both pseudo-bulk pools; Scaden with the cross-donor pool, which is standard Scaden-style training.

Architecture	Configuration	`ct1` MSE	`ct2` MSE	`ct3` MSE	avg. degr.
SimpleDNN	Full	**0.00295**	**0.00224**	**0.00305**	n/a
	Pseudo-only, donor-stable	0.00572	0.00517	0.00809	2.3×
	Pseudo-only, cross-donor	0.01722	0.00412	0.00735	3.4×
Scaden	Full	**0.00197**	**0.00151**	**0.00202**	n/a
	Pseudo-only, cross-donor	0.00282	0.00241	0.00321	1.5×

6.3 How much of the augmentation gain is donor coupling?

With the anchor established (§6.2), we turn to the augmentation beside it. Of the two ingredients on top of the anchor, the pseudo-bulk augmentation, not the gene filter (a compatibility requirement), drives most of the non-anchor gain. Why does the augmentation help? Two explanations compete. It could be donor coupling, each pseudo-bulk is drawn from a single donor's cells (*donor-stable*), which preserves that donor's real link between a cell's expression and the donor's overall composition. Or it could be simpler, that the pseudo-bulks just expose the network to a broad range of compositions. Disentangling the two, we find the gain is mostly composition coverage, with donor coupling a secondary contributor.

The comparison is the standard cross-donor recipe of deep deconvolvers like Scaden and DeSide (Menden et al., 2020; Xiong et al., 2024), which on a multi-donor reference such as ours draw cells by cell type across donors and sum them (§2). Within each of the 10 cross-validation folds (§5), instead of drawing each pseudo-bulk from one donor's cells, we merge the cells of all 27 training donors (every donor except the 3 held out in that fold, excluded to prevent leakage) into one shared pool and draw from it, so a pseudo-bulk can now mix cells from many donors. The number of pseudo-bulks (810, or 30 per donor across the 27 training donors) and their sizes match the donor-stable set exactly; the real bulks, gene filter, schedule, and seed are unchanged.

We run this on three MLPs that map the bulk to composition directly, our `SimpleDNN`, the Scaden ensemble (Menden et al., 2020), and DAISM-DNN (Lin et al., 2022) (an MLP that blends each synthetic sample with a real calibration sample), all trained on the identical pseudo-bulk pools so that only the architecture and the pool differ. All numbers are three-seed means $\{42, 7, 17\}$ (Table 4).

Table 4: Raw donor-stable vs. cross-donor mean MSE (three seeds $\{42, 7, 17\}$, everything fixed except the pseudo-bulk pool; the real bulks are in all cells). Best value per level is bold (Scaden). The raw advantage in the right column is large, but Table 5 shows most of it is composition coverage, not donor coupling.

Architecture	Pseudo-bulk pool	`ct1`	`ct2`	`ct3`	donor-stable adv.
SimpleDNN (ours)	cross-donor	0.00744	0.01393	0.01191	60–84%
	donor-stable	0.00295	0.00224	0.00305	
Scaden	cross-donor	0.00313	0.00249	0.00336	37–40%
	donor-stable	**0.00197**	**0.00151**	**0.00202**	
DAISM-DNN	cross-donor	0.00291	0.00230	0.00275	5–14%
	donor-stable	0.00255	0.00199	0.00261	

Donor-stable beats the standard recipe at every level for all three MLPs, by 60 to 84% mean MSE for our small network, 37 to 40% for Scaden, and 5 to 14% for DAISM-DNN (Table 4). That it recurs across all

three, two of which we did not design, shows the effect is not peculiar to our small network. The advantage is real, but it conflates two differences at once, the cell source (donor structure) and the breadth of the composition distribution.

Most of the gap is composition coverage, not donor identity. Uniform pooling narrows the range of cell-type compositions the pseudo-bulks cover, as a side effect. Every pooled pseudo-bulk sits near the pool-average composition, so the set clusters in a narrow band, whereas each donor-stable pseudo-bulk sits at its own donor's composition, so the set *spans* the donors' compositions. To isolate the cell-source effect from this difference in composition breadth, we add a *composition-matched* cross-donor arm. For each donor-stable pseudo-bulk we build a counterpart with the same per-type cell counts and total size but draw each type's cells from the pooled donors, so the two match in composition exactly and differ only in cell source (each matched pseudo-bulk still spans ≥ 22 donors).

Giving the cross-donor arm the same composition coverage closes most of the gap (Table 5). For `SimpleDNN` the matched arm reaches 0.0038/0.0023/0.0038 across the three levels, next to donor-stable's 0.0030/0.0022/0.0031 and far below the raw cross-donor 0.0074/0.0139/0.0119. Decomposed this way, **composition coverage accounts for most of the raw gap**, 77 to 99% for `SimpleDNN` and Scaden and 57 to 78% for DAISM-DNN, whose raw gap is near noise to begin with. The genuine donor-coupling residual, donor-stable's edge over a *composition-matched* cross-donor arm, is small: about 4 to 23% for `SimpleDNN` (largest at `ct1`/`ct3`, near zero at `ct2`), ∼5 to 13% for Scaden, and ∼2 to 4% for DAISM-DNN. Donor coupling is real and consistently positive, but it is a secondary ingredient, not the mechanism the raw gap suggested.

Table 5: Composition-matched decomposition of the raw donor-stable vs. cross-donor gap (three-seed mean MSE, all three MLPs). Table 4's two arms are donor-stable (A_1) and the raw cross-donor arm (A_2); the composition-matched arm (A_3) shares A_1's composition distribution exactly and differs only in cell source. "Composition share" is the fraction of the raw gap removed by matching composition, $(A_2-A_3)/(A_2-A_1)$. "Residual donor-coupling" and "raw adv." are both donor-stable's relative MSE advantage over a comparison arm, against the fair matched arm $(A_3-A_1)/A_3$ and against the standard cross-donor arm $(A_2-A_1)/A_2$ respectively; the latter repeats Table 4. Most of the raw gap is composition coverage, and the advantage over a fair baseline is small.

Architecture	composition share (`ct1`/`ct2`/`ct3`)	residual donor-coupling	raw adv.
SimpleDNN (ours)	81% / 99% / 91%	∼4–23%	60–84%
Scaden	91% / 77% / 86%	∼5–13%	37–40%
DAISM-DNN	78% / 74% / 57%	∼2–4%	5–14%

The residual follows the network's own defenses. The small donor-coupling residual follows the same architecture ordering as the raw gap: largest for the unconstrained `SimpleDNN`, smaller for Scaden, and near zero for DAISM-DNN. DAISM's near-zero residual has a concrete cause. Its defining step mixes every synthetic sample with a real calibration bulk ($e = f\theta + (1-f)\phi$, a real bulk θ and a synthetic mixture ϕ at a random weight f), so on average half its training signal already comes from a real anchor, the same protection our pseudo-only ablation isolates (§6.2). A network that already anchors on real bulks has little left for donor-stable sampling to add. Still, DAISM-DNN trails Scaden at every level of the headline benchmark (Table 1).

6.4 Statistical significance

The headline configuration, Scaden + ours, *significantly* beats CSCD, the strongest of the sixteen classical methods, at every granularity level (Table 6). Our deliberately small `SimpleDNN` + ours does not reach significance against CSCD ($p > 0.05$); we read this as a tie, since the point-estimate margin is small ($\sim 4 \times 10^{-4}$) and a few COPD outliers dominate the variance. The same network significantly beats the second-tier baselines PREDE.F and Bisque.F.

Table 6: Paired Wilcoxon signed-rank tests on per-sample MSE, one-sided H_1 = "ours is better"; bold p-values are significant at $\alpha = 0.05$. n counts the samples on which both the method and the comparator returned a valid prediction, falling below 30 at `ct2` because CSCD does not fit every sample (§6.1 explains why dropping these only makes the comparison more conservative).

Method (ours)	vs	Level	n	wins	mean diff	p
Scaden + ours	CSCD	`ct1`	30	18	−0.00125	**0.012**
	CSCD	`ct2`	24	22	−0.00118	$\mathbf{3.8 \times 10^{-5}}$
	CSCD	`ct3`	30	24	−0.00146	$\mathbf{1.6 \times 10^{-3}}$
SimpleDNN + ours	CSCD	`ct1`	30	14	−0.00017	0.47
	CSCD	`ct2`	24	17	−0.00047	0.057
	CSCD	`ct3`	30	18	−0.00048	0.17
	PREDE.F	`ct2`	30	29	−0.01878	$\mathbf{1.9 \times 10^{-9}}$
	Bisque.F	`ct3`	30	24	−0.00393	$\mathbf{8.5 \times 10^{-5}}$

7 Limitations and future work

Our evidence comes from a single cohort, 30 paired BAL donors from one study. The strategy is built for the paired setting, where real bulks anchor training and supply the matched ground truth that tells us whether a prediction is right. A held-out donor's real bulk is the only faithful test of whether the strategy transfers, so the meaningful way to probe generality is more paired cohorts, on other tissues and at larger scale. Such cohorts are rare, and this is one of the few that exist. That the strategy delivers this much on one small cohort is itself the argument for producing more.

8 Conclusion

We set out to improve deep cellular deconvolution by making better use of limited paired bulk + single-cell RNA-seq data. Paired pilot profiling has already been shown to improve deconvolution by providing protocol-matched in-study references and enabling paired-data-based gene filtering. This work shows that the same paired design can also be used to construct more effective training data for deep deconvolution. Instead of relying only on scarce real paired samples or only on synthetic pseudo-bulks, we combine both: pseudo-bulks generated from the matched single-cell data increase the effective number and diversity of labeled examples, while real paired bulk samples keep training calibrated to the target bulk RNA-seq distribution.

This paired-training strategy improved accuracy for every network we tested and lifted a standard Scaden-style network to the best result on a real paired BAL benchmark, ahead of every classical method and of standard pseudo-bulk deep training. Controlled experiments trace the gain to the training data rather than the architecture. Keeping the real paired bulks as anchors is the load-bearing ingredient, and the pseudo-bulk augmentation helps mainly by broadening the range of cell-type compositions the model sees. The gain is smaller for a network that already anchors on real calibration samples, consistent with this account. The broader message is that in paired deep deconvolution, how the training data are constructed can matter as much as the network it is fed to.

The practical implication is that paired pilot profiling can strengthen large bulk RNA-seq deconvolution studies without requiring single-cell profiling of the full cohort. A modest number of paired donors can provide an in-study reference, support gene-level compatibility checks, and, as shown here, generate an anchored augmented training set for deep learning. The evidence comes from a single 30-donor BAL cohort, so validation in other tissues and larger paired studies remains necessary. Nevertheless, these results support paired, bulk-anchored data augmentation as a cost-effective route for improving deep cellular deconvolution in real transcriptomic studies.

References

Hananeh Aliee and Fabian J. Theis. AutoGeneS: Automatic gene selection using multi-objective optimization for RNA-seq deconvolution. *Cell Systems*, 12(7):706–715, 2021.

Francisco Avila Cobos, José Alquicira-Hernandez, Joseph E Powell, Pieter Mestdagh, and Katleen De Preter. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nature communications*, 11 (1):5650, 2020.

Maayan Baron, Adrian Veres, Samuel L Wolock, Aubrey L Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K Wagner, Shai S Shen-Orr, Allon M Klein, Douglas A Melton, and Itai Yanai. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Systems*, 3(4):346–360, 2016. doi: 10.1016/j.cels.2016.08.011.

Yanshuo Chen, Yixuan Wang, Yuelong Chen, Yuqi Cheng, Yumeng Wei, Yunxiang Li, Jiuming Wang, Yingying Wei, Ting-Fung Chan, and Yu Li. Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. *Nature Communications*, 13(1):6735, 2022. doi: 10.1038/s41467-022-34550-9.

Tinyi Chu, Zhong Wang, Dana Pe'er, and Charles G. Danko. Cell type and gene expression deconvolution with BayesPrism enables bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nature Cancer*, 3(4):505–517, 2022.

Natalie R. Davidson, Fan Zhang, and Casey S. Greene. BuDDI: Bulk deconvolution with domain invariance to predict cell-type-specific perturbations from bulk. *PLoS Computational Biology*, 21(1):e1012742, 2025. doi: 10.1371/journal.pcbi.1012742.

Meichen Dong, Aatish Thennavan, Eugene Urrutia, Yun Li, Charles M. Perou, Fei Zou, and Yuchao Jiang. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Briefings in Bioinformatics*, 22(1):416–427, 2021.

Ting Gong and Joseph D. Szustakowski. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics*, 29(8):1083–1085, 2013.

Simon Grouard, Khalil Ouardini, Yann Rodriguez, Jean-Philippe Vert, and Almudena Espin-Perez. Joint probabilistic modeling of pseudobulk and single-cell transcriptomics enables accurate estimation of cell type composition. In *ICML 2025 Workshop on Generative AI for Biology*, 2025.

Yuning Hao, Ming Yan, Blake R. Heath, Yu L. Lei, and Yuying Xie. Fast and robust deconvolution of tumor infiltrating lymphocyte from expression profiles using least trimmed squares. *PLoS Computational Biology*, 15(5):e1006976, 2019.

Ariel A Hippen, Dalia K Omran, Lukas M Weber, Euihye Jung, Ronny Drapkin, Jennifer A Doherty, Stephanie C Hicks, and Casey S Greene. Performance of computational algorithms to deconvolve heterogeneous bulk ovarian tumor tissue depends on experimental factors. *Genome Biology*, 24(1):239, 2023. doi: 10.1186/s13059-023-03077-7.

Mengying Hu and Maria Chikina. Heterogeneous pseudobulk simulation enables realistic benchmarking of cell-type deconvolution methods. *Genome Biology*, 25(1):169, 2024. doi: 10.1186/s13059-024-03292-w.

Yushan Hu, Ziying Liu, Danielle Tsao, Janice M. Leung, Firoozeh V. Gerayeli, Xuan Li, Xiaojian Shao, Don D. Sin, and Xuekui Zhang. Real-paired single-cell/bulk rna-seq benchmark and a practical protocol for accurate cell-type deconvolution in human bal samples. *bioRxiv*, 2026. doi: 10.64898/2026.01.14.699304. URL `https://www.biorxiv.org/content/10.64898/2026.01.14.699304v1`. Preprint; the benchmark this work compares against (sc.ubchli reference, 10-fold leave-3-out CV).

Gregory J Hunt, Saskia Freytag, Melanie Bahlo, and Johann A Gagnon-Bartsch. Dtangle: accurate and robust cell type deconvolution. *Bioinformatics*, 35(12):2093–2099, 2019.

Brandon Jew, Marcus Alvarez, Elior Rahmani, Zong Miao, Arthur Ko, Kristina M Garske, Jae Hoon Sul, Kirsi H Pietiläinen, Päivi Pajukanta, and Eran Halperin. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Nature communications*, 11(1):1971, 2020.

Kai Kang, Qian Meng, Igor Shats, David M. Umbach, Melissa Li, et al. CDSeq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLoS Computational Biology*, 15(12):e1007510, 2019.

Ali Karimnezhad. More accurate estimation of cell composition in bulk expression through robust integration of single-cell information. *Bioinformatics Advances*, 2(1):vbac049, 2022. doi: 10.1093/bioadv/vbac049.

Robin Khatri, Pierre Machart, and Stefan Bonn. Dissect: deep semi-supervised consistency regularization for accurate cell type fraction and gene expression estimation. *Genome Biology*, 25(1):112, 2024. doi: 10.1186/s13059-024-03251-5.

Taiwen Li, Jingxin Fu, Zexian Zeng, David Cohen, Jing Li, Qianming Chen, Bo Li, and X Shirley Liu. Timer2.0 for analysis of tumor-infiltrating immune cells. *Nucleic acids research*, 48(W1):W509–W514, 2020.

Ziyi Li and Hao Wu. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biology*, 20(1):190, 2019.

Yating Lin, Haojun Li, Xu Xiao, Lei Zhang, Kejia Wang, Jingbo Zhao, Minshu Wang, Frank Zheng, Minwei Zhang, Wenxian Yang, Jiahuai Han, and Rongshan Yu. Daism-dnnxmbd: Highly accurate cell type proportion estimation with in silico data augmentation and deep neural networks. *Patterns*, 3(3):100440, 2022. doi: 10.1016/j.patter.2022.100440.

Alba Lomas Redondo, Jose M Sánchez Velázquez, Álvaro J García Tejedor, and Víctor Javier Sánchez-Arévalo Lobo. Deep learning based deconvolution methods: a systematic review. *Computational and Structural Biotechnology Journal*, 27:2544–2565, 2025. doi: 10.1016/j.csbj.2025.05.038.

Kevin Menden, Matteo Marouf, Sergio Oller, Anupriya Dalmia, Daniel Sumner Magruder, Karin Kloiber, Peter Heutink, and Stefan Bonn. Deep learning–based cell composition analysis from tissue expression profiles. *Science advances*, 6(30):eaba2619, 2020.

Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature methods*, 12(5):453–457, 2015.

Aaron M Newman, Chloé B Steen, Chih Long Liu, Andrew J Gentles, Aadel A Chaudhuri, Florian Scherer, Michael S Khodadoust, Mohammad S Esfahani, Bogdan A Luca, David Steiner, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature biotechnology*, 37(7): 773–782, 2019.

Yufang Qin, Weiwei Zhang, Xiaoqiang Sun, Siwei Nan, Nana Wei, Hua-Jun Wu, and Xiaoqi Zheng. Deconvolution of heterogeneous tumor samples using partial reference signals. *PLoS Computational Biology*, 16 (11):e1008452, 2020. doi: 10.1371/journal.pcbi.1008452.

Julien Racle, Kaat De Jonge, Petra Baumgaertner, Daniel E Speiser, and David Gfeller. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *elife*, 6:e26476, 2017.

Wenjun Shen, Yunfei Hu, Yuanfang Lei, Hau-San Wong, Cheng Liu, Si Wu, and Xin Maizie Zhou. CSsingle: a unified tool for robust decomposition of bulk and spatial transcriptomic data across diverse single-cell references. *Nucleic Acids Research*, 54(8):gkag410, 2026. doi: 10.1093/nar/gkag410.

Carlos Torroja and Fatima Sanchez-Cabo. Digitaldlsorter: deep-learning on scRNA-seq to deconvolute gene expression data. *Frontiers in Genetics*, 10:978, 2019.

Daphne Tsoucas, Rui Dong, Haide Chen, Qian Zhu, Guoji Guo, and Guo-Cheng Yuan. Accurate estimation of cell-type composition from gene expression data. *Nature Communications*, 10(1):2975, 2019.

Xuran Wang, Jihwan Park, Katalin Susztak, Nancy R Zhang, and Mingyao Li. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature communications*, 10(1):380, 2019.

Xin Xiong, Yerong Liu, Dandan Pu, Zhu Yang, Zedong Bi, Liang Tian, and Xuefei Li. DeSide: A unified deep learning approach for cellular deconvolution of tumor microenvironment. *Proceedings of the National Academy of Sciences*, 121(46):e2407096121, 2024.

Aleksandr Zaitsev et al. Precise reconstruction of the TME using bulk RNA-seq and a machine learning algorithm trained on artificial transcriptomes. *Cancer Cell*, 40(8):879–894.e16, 2022. doi: 10.1016/j.ccell.2022.07.006.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

A Full per-sample distributions across all methods

Table 1 reports mean metrics for the top methods. Figure 1 shows the full picture behind those means, the per-sample distribution of every metric, for all eighteen methods, at all three granularity levels (methods and metrics are defined in §5). Three patterns appear that the means alone cannot show. First, both networks trained with our strategy sit at the top of every one of the nine panels, so their advantage holds on MSE, R^2, and Cor alike, not only on the metric Table 1 sorts by. Second, the deep configurations have far tighter distributions than the classical methods, with accuracy that is consistent across donors, whereas several classical methods swing widely from sample to sample. Third, a few classical methods fail outright on the harder per-sample metrics, visible as the long negative R^2 tails. An R^2 below zero means a prediction worse than simply guessing each sample's mean composition.

B Supporting analyses for the composition account

Two analyses corroborate §6.3's composition account of the raw cross-donor deficit from two angles, why it is removable (and why the canonical method needs no mixup or KL) and where it concentrates. Both use our `SimpleDNN`.

The deficit is a fixable distributional artifact. If the raw cross-donor pool is weak mainly because its narrow, pool-average compositions create a train/test mismatch, then a regularizer that broadens or smooths the training signal should partly rescue it. It does. In a 2×2 over (pseudo-bulk pool) $\times$ (training-time regularization) (Table 7), mixup/KL rescues the cross-donor pool substantially (0.0074/0.0139/0.0119 $\rightarrow$ 0.0035/0.0025/0.0033) while barely moving donor-stable, which needs no such help. The raw cross-donor deficit is largely a fixable distributional artifact, and the canonical method needs no mixup or KL because composition-diverse augmentation already avoids the mismatch at the source (cf. §4).

Table 7: Mean MSE (three seeds $\{42, 7, 17\}$) under the 2×2 interaction of pseudo-bulk pool (donor-stable vs. cross-donor) and training-time regularization (none vs. +mixup+KL), our `SimpleDNN`. Donor-stable wins in both rows; cross-donor depends on the regularizer that donor-stable does not need.

Pseudo-bulk pool	Regularization	`ct1`	`ct2`	`ct3`
donor-stable	none (canonical)	0.0029	0.0022	0.0030
	+mixup+KL	0.0030	0.0022	0.0029
cross-donor	none	0.0074	0.0139	0.0119
	+mixup+KL	0.0035	0.0025	0.0033

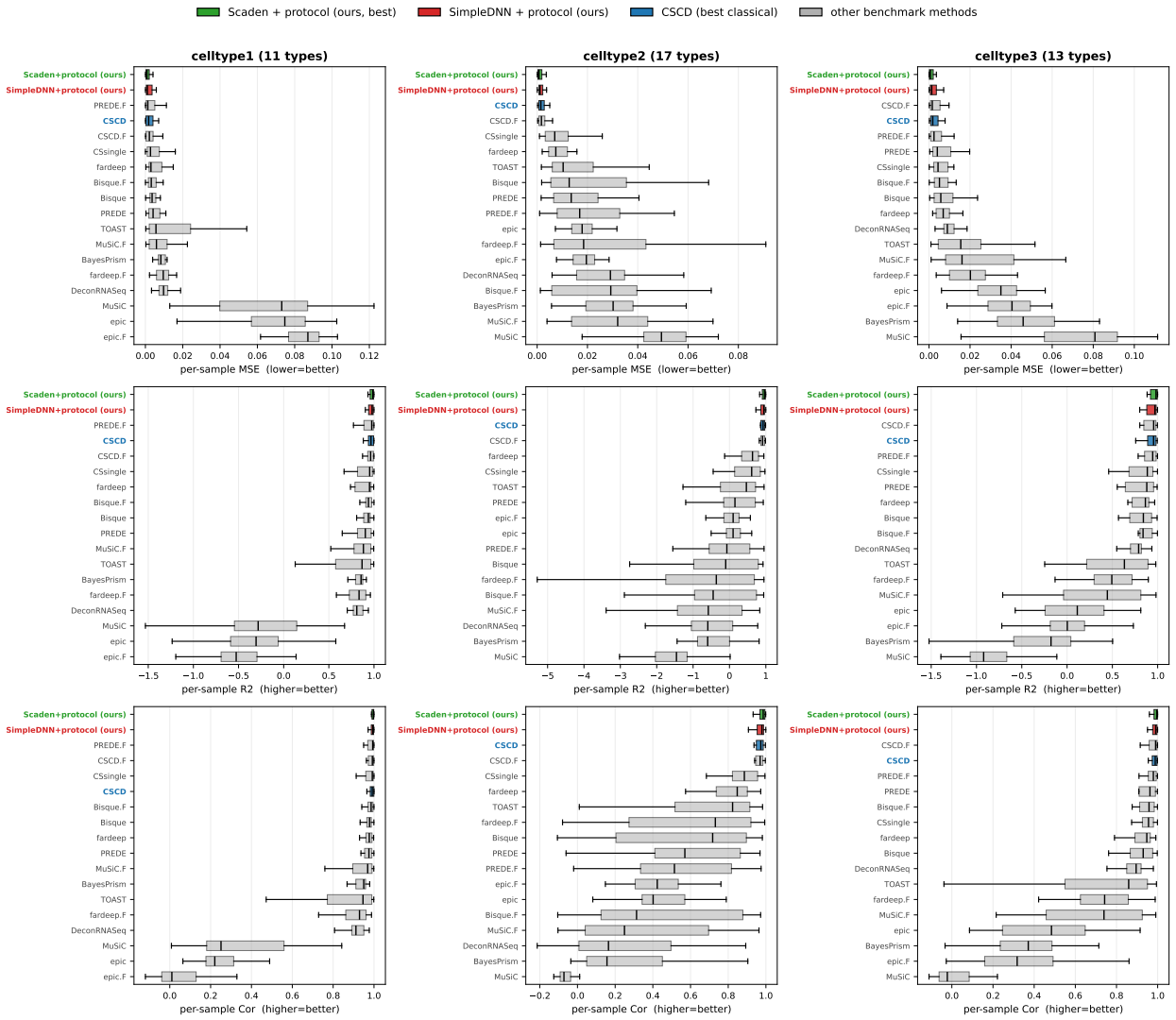

Figure 1: Per-sample deconvolution accuracy on the `sc.ubchli` reference (30 held-out samples, 10-fold CV) for all eighteen methods. Rows are MSE (lower better), R^2, and Cor (higher better); columns are the three granularity levels (`ct1` / `ct2` / `ct3`). Green and red are Scaden and `SimpleDNN` trained with our strategy; blue is the strongest classical method (CSCD); gray are the other fifteen. Methods are sorted by median *independently within each panel*, so a method's vertical position is not comparable across panels. Methods and metrics are defined in §5.

The deficit lives in the macrophage subtypes. Splitting the raw cross-donor excess error across cell types (`ct2`, three-seed mean; Table 8) localizes almost all of it, about 90%, in the macrophage and monocyte subpopulations, and 87% of the total in just two of them (`Macrophage_CCL18` and `Alveolar_Macrophage`). This fits the composition account. Macrophage-subtype proportions vary widely from donor to donor, so the pool-average cross-donor arm sits at one central composition and covers that spread poorly, never showing the network the subtype mixtures the real test bulks present. Composition-diverse augmentation, whether donor-stable or composition-matched, restores that coverage. The decomposition is correlational, not a causal test.

Table 8: The cross-donor penalty is concentrated in macrophage subtypes. Per-class share of the donor-stable-vs-cross-donor excess MSE at `ct2` (cross-donor minus donor-stable per-class MSE, as a fraction of the total gap $0.0139 - 0.0022 = 0.0117$; our `SimpleDNN`, mean over three seeds).

Cell type	share of excess MSE
`Macrophage_CCL18`	59.2%
`Alveolar_Macrophage`	27.9%
other macrophage/monocyte (5 subtypes)	2.4%
Macrophage/monocyte subtotal	**89.5%**
`CD4_Tcell`	5.5%
other 9 cell types	5.0%

Broader Impact Statement

This work may make deep cellular deconvolution more practical for biomedical studies where bulk RNA-seq is available at scale but single-cell profiling is feasible only for a limited subset of samples. By using a small paired bulk and single-cell RNA-seq pilot cohort to construct an augmented, bulk-anchored training set, the proposed strategy can improve cell-type composition estimation without requiring single-cell sequencing of every donor.

This can support more informative reuse of large bulk RNA-seq cohorts and help researchers study disease-associated changes in cellular composition in heterogeneous tissues. The approach also promotes a cost-effective study design in which limited paired profiling strengthens downstream analysis of much larger bulk transcriptomic datasets.