DISENTANGLING MARGINAL CAPACITY FROM GEOMETRY IN OPTIMAL-TRANSPORT LABEL TRANSFER

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimal transport (OT) is widely used to align distributions and transfer labels between unpaired source and target data, but source–target composition shifts can cause rare classes to be systematically under-predicted even with high transport confidence. We identify a source of this failure in the transport marginal: under the standard uniform weighting of source points, each class has transport capacity proportional to its source frequency, which can be insufficient to meet its target demand even when separated well in the embedding space. We show that displaced mass is preferentially absorbed by high-capacity classes rather than geometrically similar classes, producing confident errors that are difficult to detect from the transport plan. Downstream read-outs, entropic smoothing, and marginal relaxation provide limited recovery, whereas matching the transport marginal to the target composition substantially improves label transfer. On single-cell data, the oracle composition-matched marginal raises macro-F1 from 0.322 to 0.623, and the best deployable estimator closes 69% of the oracle gap. Correcting the marginal also separates a coverage-dependent component of error from residual error associated with class separability, showing that transport capacity and representation geometry are distinct sources of transfer error.

1 INTRODUCTION

Optimal transport (OT) is widely used to transfer labels between unpaired source and target data: given a shared representation, target labels are inferred from the source mass transported to each target point. This formulation appears across domain adaptation (Courty et al., 2017; Fatras et al., 2021) and cross-domain alignment (Cao et al., 2022; Cao & Gao, 2022; Klein et al., 2025; Bunne et al., 2024), where the transport coupling provides a common mechanism for matching source and target distributions. Yet when source and target compositions differ, rare classes can transfer poorly despite high confidence, with errors surfacing as false positives on abundant classes (Fig. 1). This pattern is often attributed to representation quality (Tejada-Lapuerta et al., 2025; Cui et al., 2024; Kedzierska et al., 2025), but can the transport objective itself create this asymmetry?

We study this question through the source marginal of the transport plan. The marginal determines how much transport mass each source class can supply, and under the standard uniform weighting of source points (Cuturi, 2013; Peyré & Cuturi, 2019), this capacity is proportional to source class frequency. When the target composition differs from the source composition, a class may therefore have less transport mass available than the target mass it needs, *even when the embedding separates it well*.

This failure is difficult to detect because rare classes can be displaced into large, well-supported classes with high transport confidence (Fig. 1b). Because the displaced mass can still receive high-confidence labels, output-side statistics can obscure rather than reveal the underlying capacity failure (Sec. 3.2). Downstream

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

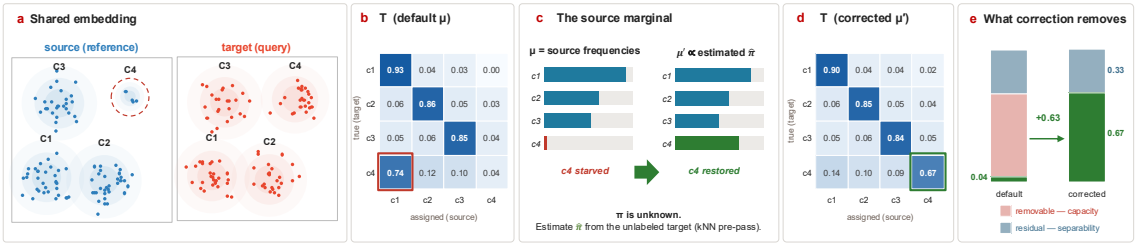


Figure 1: **Capacity and correction. (a)** c_4 is rare in the source yet well separated; c_1, c_2 overlap. **(b)** Under the default $\mu =$ source frequencies, c_4 is starved and its mass is absorbed by the abundant c_1. **(c)** Replacing it with $\mu' = \hat{\pi}/n^s$, $\hat{\pi}$ estimating the target composition, restores c_4's budget. **(d)** Correct assignments return; the $c_1 \leftrightarrow c_2$ off-diagonal is unchanged. **(e)** Recovered mass is the capacity term; the residual tracks separability.

modifications of the transport plan are similarly limited: alternative read-outs (El Hamri et al., 2021), entropic smoothing (Cuturi, 2013), and relaxing the marginal constraint recover only a fraction of the lost rare-class performance, while post-hoc class-prior reweighting recovers substantially less than correcting the marginal itself (Sec. 3.3). This limitation is a consequence of where the failure occurs: it arises in the transport coupling, before the final label read-out.

We therefore reshape the transport marginal to reflect the target composition rather than the source frequencies. In practice, however, the target composition is unknown, and we test whether it can be estimated from unlabeled target data well enough to recover the lost rare-class performance (Fig. 1c, Sec. 4). We evaluate this composition-matched marginal when source and target compositions differ under two complementary domain shifts: cross-modal single-cell transfer (Janesick et al., 2023; Li et al., 2022), where the measurement process changes, and image domain adaptation (Venkateswara et al., 2017), where the shift is primarily stylistic and coverage can be controlled independently of class separability. The oracle composition-matched marginal raises macro-F1 for OT label transfer from 0.322 to 0.623 on single-cell data, while the best deployable estimator closes 69% of the oracle gap on single-cell data and 59% on Office-Home (Sec. 4). We also find that the correction should be applied selectively: its benefit is concentrated in capacity-limited classes and can become negative when coverage is sufficient (Sec. 4). Together, these results show that correcting the transport marginal removes a coverage-dependent component of label-transfer error and exposes a residual component associated with the representation.

To summarize, we make the following contributions:

- **Capacity-driven failure.** We identify marginal capacity as a source of rare-class transfer errors, with displaced mass flowing to high-capacity rather than geometrically similar classes.
- **Upstream failure.** We show that these errors are difficult to detect and that downstream read-outs, smoothing, and marginal relaxation provide limited recovery.
- **Marginal correction.** Matching the marginal to the target composition substantially improves transfer, but only for capacity-limited classes; its practical use depends on estimating the target composition.
- **Capacity–geometry separation.** Correcting the marginal removes the coverage-dependent component of error, revealing residual error associated with class separability.

2 OPTIMAL-TRANSPORT LABEL TRANSFER AND MARGINAL CAPACITY

Optimal-transport label transfer Let $\mathcal{D}^{\mathrm{s}} = \{(x_j^{\mathrm{s}}, y_j)\}_{j=1}^{n^{\mathrm{s}}}$ be a labelled source dataset of n^{s} points, with $x_j^{\mathrm{s}} \in \mathcal{X}^{\mathrm{s}}$ and class labels $y_j \in \{1, \dots, K\}$, and let $\mathcal{D}^{\mathrm{t}} = \{x_i^{\mathrm{t}}\}_{i=1}^{n^{\mathrm{t}}}$ be an unlabelled target dataset of n^{t} points, with $x_i^{\mathrm{t}} \in \mathcal{X}^{\mathrm{t}}$. A frozen encoder $\phi\colon \mathcal{X}^{\mathrm{s}} \cup \mathcal{X}^{\mathrm{t}} \to \mathbb{R}^d$ maps source and target data into a shared d-dimensional space. We use the squared Euclidean distance there as the transport cost (Peyré & Cuturi, 2019; Courty et al., 2017),

$$C_{ij} = \left\|\phi(x_i^{\mathrm{t}}) - \phi(x_j^{\mathrm{s}})\right\|^2.$$

Let $\mu \in \Delta^{n^{\mathrm{s}}}$ and $\nu \in \Delta^{n^{\mathrm{t}}}$ denote the source and target marginals, where Δ^n is the probability simplex in $\mathbb{R}^n$. We use a uniform target marginal throughout, $\nu_i = 1/n^{\mathrm{t}}$. The transport plan $T \in \mathbb{R}_+^{n^{\mathrm{t}} \times n^{\mathrm{s}}}$ is obtained by solving entropic unbalanced OT (UOT) (Cuturi, 2013; Chizat et al., 2018; Séjourné et al., 2023):

$$T^\star = \operatorname*{arg\,min}_{T \geq 0} \langle C, T\rangle + \varepsilon H(T) + \tau\left[\mathrm{KL}(T\mathbf{1} \,\|\, \nu) + \mathrm{KL}\big(T^\top \mathbf{1} \,\big\|\, \mu\big)\right]. \tag{1}$$

Here $\langle C, T\rangle = \sum_{ij} C_{ij} T_{ij}$, $H(T) = \sum_{ij} T_{ij}(\log T_{ij} - 1)$, and $\mathrm{KL}(a \,\|\, b) = \sum_\ell [a_\ell \log(a_\ell/b_\ell) - a_\ell + b_\ell]$ is the generalized Kullback–Leibler divergence. The parameter ε controls the entropic regularization, while τ controls how strongly the source and target marginals are enforced; as $\tau \to \infty$, the solution approaches balanced OT.

Marginal capacity The source marginal μ determines the nominal amount of transport mass available from each source class. We refer to this class-level mass as its *marginal capacity*: class k has marginal capacity $\sum_{j:y_j=k} \mu_j$. Under the standard choice $\mu_j = 1/n^{\mathrm{s}}$ (Cuturi, 2013; Peyré & Cuturi, 2019), this becomes $n_k^{\mathrm{s}}/n^{\mathrm{s}}$, where n_k^{s} is the number of source points of class k. Thus, marginal capacity is proportional to source class frequency.

Read-outs The transport plan is converted into target labels by aggregating transported mass from source points of each class. The standard read-out is

$$\hat{y}_i = \operatorname*{arg\,max}_k \sum_{j\colon y_j = k} T_{ij}^\star.$$

Other read-outs use the same class-aggregated mass, including direct or tempered posterior probabilities and label propagation (El Hamri et al., 2021). Because these read-outs operate on a fixed transport plan, they can be varied independently of the transport solve; we use this separation in Sec. 3.1.

Coverage ratio For each class k, let n_k^{t} denote the number of target points of that class and let $\pi_k = n_k^{\mathrm{t}}/n^{\mathrm{t}}$ denote its target composition. We measure source supply relative to target demand using the coverage ratio

$$r_k = \frac{n_k^{\mathrm{s}}}{n_k^{\mathrm{t}}},$$

the number of source points of class k available per target point of that class. Under uniform ν, the target marginal assigns total mass π_k to the target points of class k.

3 THE CAPACITY FAILURE

We next test whether the failure predicted by the source marginal is driven by transport capacity rather than by the representation or the transport solver. We first characterize its signature, then ask where the displaced mass goes, and finally test which components of the OT formulation can remove it.

Data We evaluate the capacity hypothesis in two complementary domain-shift settings. In single-cell cross-modal transfer, the source is dissociated RNA sequencing and the target is spatial transcriptomics from the same tissue, so the domain gap is primarily a change in measurement process (Li et al., 2022). We use Jint and Janesick (Janesick et al., 2023), together with SpatialCorpus (Tejada-Lapuerta et al., 2025) as a no-composition-shift control. In image domain adaptation, we use Office-Home (Venkateswara et al., 2017), where the domain gap is primarily stylistic; because the dataset is class balanced, we independently vary class coverage to separate it from representation separability. Full experimental details are given in App. C.

Evaluation We report macro recall, macro-F1, and micro recall. An intervention counts as a fix only if it improves macro-F1 without reducing micro recall. Results are stratified by capacity and evaluated at the target's natural composition unless stated otherwise. Full metric definitions and protocol are given in App. C.

3.1 CAPACITY SIGNATURE

Under the default marginal on single-cell data, the largest class reaches recall 0.95 but precision only 0.19, while a capacity-limited class reaches recall 0.05 with precision 1.00 as can be seen in Table 1. The two errors are asymmetric: the largest class absorbs false positives, while the capacity-limited class is rarely predicted but, when predicted, is usually correct. This is consistent with a class being short of transport mass rather than being confused with another class.

Table 1: Per-class recall / precision on single-cell data (Jint), natural target composition, 3 seeds. The default marginal produces the capacity signature: a capacity-limited class has low recall but high precision, while the largest class has high recall but low precision.

Source marginal	B_Plasma	T_NK	Myeloid	Malignant	Epithelial	Endothelial	Stromal
source frequencies	.65/.63	**.95/.19**	.30/.53	.64/.55	.36/.53	**.05/1.00**	.09/.89
uniform per class	.73/.66	.80/.51	.43/.69	.66/.63	.72/**.26**	.64/**.29**	.60/.73
composition-matched	.78/.68	**.75/.71**	.61/.62	.61/.78	.45/.48	.46/.72	**.69/.72**

The pattern is not specific to our implementation or read-out. Three independent entropic-OT implementations produce recall between 0.009 and 0.020 for the same capacity limited class on identical embeddings, points, labels, folds, and seeds, whereas a kNN vote on the same embedding reaches 0.295 (Table 13). Thus, the failure persists across OT solvers even when the underlying representation is fixed.

3.2 WHERE THE DISPLACED MASS GOES

The capacity signature could still reflect poor geometry: perhaps starved classes are poorly represented, and their mass simply flows to nearby classes. We therefore compare two explanations for the failure: limited representation geometry and limited transport capacity. If geometry drives the failure, mass should accumulate in classes that are close in the embedding. If capacity drives the failure, mass should instead accumulate in classes with more available source mass.

We test these two possibilities for each below-capacity class t and candidate absorber c by relating the fraction of t's target mass assigned to c to both the source mass available to c and the embedding distance between t and c, while controlling for the other. Because capacity and embedding distance are correlated in the data, we use partial Spearman correlations to measure the association of each with absorbed mass while controlling for the other (Table 2).

Table 2: Where displaced mass goes. Partial Spearman correlations of absorbed share with class capacity and embedding distance, each controlling for the other.

	single-cell (Jint)	Office-Home
partial ρ(absorbed share, **capacity** | distance)	$\mathbf{+0.566}$	$\mathbf{+0.329}$
partial ρ(absorbed share, distance | capacity)	$+0.144$	-0.206
n (class, absorber) pairs / units	198 / 27	14,080 / 12

Capacity is positively associated with absorbed mass in both domains: after controlling for distance, classes with more available source mass absorb more of the displaced mass. The association with distance is small and inconsistent across domains. The capacity partial is $+0.566$ in single-cell data and $+0.329$ on Office-Home, compared with distance partials of $+0.144$ and -0.206, respectively (Table 2). Thus, displaced mass follows available capacity rather than embedding proximity.

This result is consistent with the categorical analysis: in 30 of 33 below-capacity observations, the absorbing class is the largest source class, while it is never the nearest class. Repeating the analysis with a PCA-based representation gives the same result (Table 14, App. G). The failure therefore does not disappear when only the representation is improved: the displaced mass is allocated according to the source marginal rather than embedding proximity. App. D shows the same effect in the coupling itself on a single fold.

3.3 Detection and interventions

The failure is also difficult to detect or repair after the transport plan has been computed: displaced points are often assigned with high confidence to well-supported classes. We therefore test which components of the OT formulation can recover the missing rare-class performance by varying the entropic regularization ε, the marginal penalty τ, and the source marginal μ under matched conditions. We define a class as below capacity when $r_k < 0.15$. We fixed this threshold before the marginal experiments using a preliminary fit of the saturating recall-versus-coverage law of App. B on held-out single-cell data; the ordering of the main comparisons is unchanged at thresholds of 0.05, 0.10, and 0.30 (Table 14). For the comparison, we use four Office-Home domain pairs and three seeds, yielding 235 below-capacity and 545 above-capacity class-observations.

Table 3: Effect of the three OT parameters in Eq. 1, stratified by capacity. Entries are changes in macro recall relative to the default source marginal on Office-Home.

intervention	$r_k < 0.15$	$r_k \geq 0.15$
marginal shape (matched to target)	$\mathbf{+0.366}$	$+0.016$
marginal shape (uniform)	$+0.274$	$+0.038$
$\tau = 0.01$ (most relaxed)	$\mathbf{+0.208}$	$+0.053$
$\tau = \infty$ (balanced Sinkhorn)	-0.021	-0.013
ε ($3\times$ default)	$\mathbf{+0.000}$	-0.012
permuted marginal (control)	-0.015	-0.079

The errors are difficult to detect from the transport plan itself. Thirty of 33 below-capacity observations are confidently wrong, retaining 85% of normal posterior confidence, and five output-side detectors identify only 10 of 135 unit evaluations under our fixed detection criterion. Transport entropy is particularly misleading, because displaced points land confidently inside large, well-supported classes. Varying the entropic regularization does not rescue detection either: across seven values, the best detector's selectivity increases from 2.94 to 23.33 while below-capacity recall falls from 0.026 to 0.000 (Table 11, App. E).

Downstream changes to the transport plan also provide limited recovery. Across ten read-outs of a single converged plan, below-capacity recall increases only from 0.040 to at most 0.063 among read-outs based on transported mass (Table 8). Across four Office-Home domain pairs and three seeds, changing the entropic regularization produces essentially no gain, while relaxing the marginal penalty provides only partial recovery. Matching the marginal to the target composition is the most effective intervention below capacity, improving macro recall by $+0.366$, compared with $+0.208$ for the most relaxed marginal penalty (Table 3); above capacity, the effects are small. Relaxing τ allows the plan to deviate from the source marginal, but it does not determine where the released mass should go, and combining τ tuning with marginal correction provides no significant additional gain (Table 10). A permutation control shows that the improvement is not due to changing the marginal arbitrarily: keeping the corrected mass values but assigning them to the wrong classes performs worse than the default marginal on both domains (Table 3).

These results localize the failure to the class-level mass allocation: the relevant intervention is to assign mass to the classes that need it, rather than to change the read-out or relax the OT regularization.

4 Correcting the marginal

If marginal capacity causes the failure, correcting the source marginal should recover the missing rare-class performance. We first test this with the true target composition, then ask what remains after the correction, and finally evaluate whether the target composition can be estimated well enough for the correction to be used in practice. We use the same datasets and evaluation protocol introduced in Sec. 3.

4.1 What the correction repairs

To correct the capacity mismatch, we assign source mass to each class in proportion to its target composition rather than its source frequency:

$$\mu_j = \frac{\pi_{y_j}}{n^{\mathrm{s}}_{y_j}}, \qquad \pi_k = \frac{n^{\mathrm{t}}_k}{n^{\mathrm{t}}}. \tag{2}$$

We first evaluate this marginal using the oracle target composition π. This oracle setting isolates the effect of correcting the marginal and provides an upper bound on what a marginal-based correction can achieve.

The correction substantially improves label transfer. On single-cell data, macro-F1 increases from 0.322 to 0.623, while micro recall increases from 0.393 to 0.718. For the single-cell setting, we use three folds and three seeds; for Office-Home, we use four domain pairs and three seeds (Table 4). Uniformly balancing the source mass across classes does not provide the same correction: it assumes a uniform target composition and can improve rare-class recall by sacrificing the dominant class. On one Janesick replicate, uniform balancing increases macro recall while reducing micro recall from 0.477 to 0.346, with the largest target class falling from 0.472 to 0.069 (Table 17, App. G).

The benefit depends on coverage. It decreases monotonically with the coverage ratio and changes sign above capacity (Fig. 8, App. G). Across Office-Home, the correction ranges from $+0.473$ to -0.178 across coverage deciles, while the single-cell data range from $+0.531$ to -0.355 across five coverage bins. The

Table 4: Label-transfer performance under different source marginals and target-composition estimates. Results are shown for single-cell data (Jint) and Office-Home under the natural target composition.

	single-cell (Jint)				Office-Home			
estimate of π	macro F1		micro	below	macro F1		micro	%
source frequencies	0.433	0.322	0.393	0.030	0.392	0.359	0.437	n/a
permuted (control)	n/a	n/a	n/a	n/a	0.333	0.284	0.360	n/a
uniform	0.654	0.512	0.557	0.533	0.501	0.396	0.448	6%
kNN pre-pass	0.648	0.556	0.616	0.463	0.487	0.428	0.518	48%
JCPOT	0.585	0.505	0.595	0.280	0.467	0.434	0.537	**59%**
BBSE	0.528	0.439	0.451	0.371	0.490	0.421	0.486	29%
RLLS	0.528	0.439	0.451	0.371	0.493	0.423	0.490	31%
MLLS	0.516	0.417	0.409	**0.672**	0.494	0.433	0.510	43%
oracle π	0.619	**0.623**	**0.718**	0.546	0.514	**0.515**	**0.606**	n/a

crossing point is dataset-dependent, so we do not treat the capacity threshold as a universal break-even point. The same qualitative trend is obtained with a deployable estimate of the target composition, with smaller gains and a later break-even point (Fig. 6, App. F).

4.2 Separating capacity from geometry

Correcting the marginal removes the capacity constraint, so the remaining error should reflect limitations of the representation. We test this by decomposing each class's error into a *removable* component, the gain from the composition-matched marginal, and a *residual* component, the error that remains after correction.

We use Office-Home for this analysis because its class-balanced design lets us vary coverage independently of representation separability. Across 780 class observations, the removable component is strongly associated with coverage ($\rho = -0.809$) but not with separability ($\rho = -0.009$), whereas the residual component is associated with separability ($\rho = -0.337$) but not with coverage ($\rho = -0.005$) (Fig. 7, App. G).

This double dissociation shows that correcting the marginal removes a coverage-dependent component of transfer error while leaving a residual component associated with class separability. The residual is therefore not simply a consequence of class rarity: some well-covered classes remain poorly transferred when their representation is not separable.

The decomposition is defined using the oracle target composition. Repeating the analysis with estimated target compositions preserves the same qualitative asymmetry across estimators, although the exact dissociation is weaker (Table 12, App. F).

4.3 Estimating the target composition

The composition-matched marginal requires the target composition π, which is unknown in practice. We therefore compare estimators of π by feeding each estimate into the same marginal construction, so that differences in performance reflect the composition estimate rather than the transport procedure. We consider three label-shift estimators, BBSE (Lipton et al., 2018), RLLS (Azizzadenesheli et al., 2019), and MLLS/EM (Saerens et al., 2002; Alexandari et al., 2020), and two transport-native estimators, JCPOT (Redko et al., 2019) and a kNN pre-pass.

The label-shift estimators perform similarly and recover little of the oracle gain under the measurement-process shift. On single-cell data, BBSE, RLLS, and MLLS close only 4.7–17.8% of the $+0.324$ micro-

recall gap to the oracle, compared with 29–43% on Office-Home. Their similar performance across three datasets indicates an assumption-level limitation rather than an optimization issue: these methods assume that $p(x \mid y)$ is invariant across domains, an assumption violated by a change in measurement process.

Transport-native estimators perform better on both domains, although the best estimator depends on the shift. The kNN pre-pass closes 69% of the oracle gap on single-cell data, while JCPOT closes 59% on Office-Home. No estimator is consistently best across the two domain shifts. The correction also does not require an exact composition estimate: an estimate only 30% of the way from the source composition toward the target already closes half of the oracle gain (Sec. 4.4).

4.4 How close must the composition estimate be?

Section 4.3 shows that practical estimators of the target composition are imperfect, yet can recover a substantial fraction of the oracle correction. We therefore ask how much of the source-to-target composition shift must be recovered for the correction to be effective. We interpolate between the source and target compositions,

$$\pi_\alpha = (1-\alpha)\pi^{\mathrm{s}} + \alpha\pi, \qquad \mu_j = \frac{\pi_\alpha[y_j]}{n^{\mathrm{s}}_{y_j}},$$

where $\alpha = 0$ gives the default source marginal and $\alpha = 1$ gives the oracle marginal. We then evaluate the same transport procedure along this path, keeping the remaining OT parameters fixed.

Macro-F1 increases rapidly as α moves from the source composition toward the target composition, and then increases more slowly (Fig. 2, 3). Both domains recover half of the total oracle gain by $\alpha = 0.30$; 90% of the gain is reached by $\alpha = 0.70$ on Office-Home and $\alpha = 0.80$ on single-cell data. Thus, recovering roughly one third of the source-to-target composition shift already captures half of the available macro-F1 gain. The target composition therefore does not need to be estimated exactly for the marginal correction to be effective.

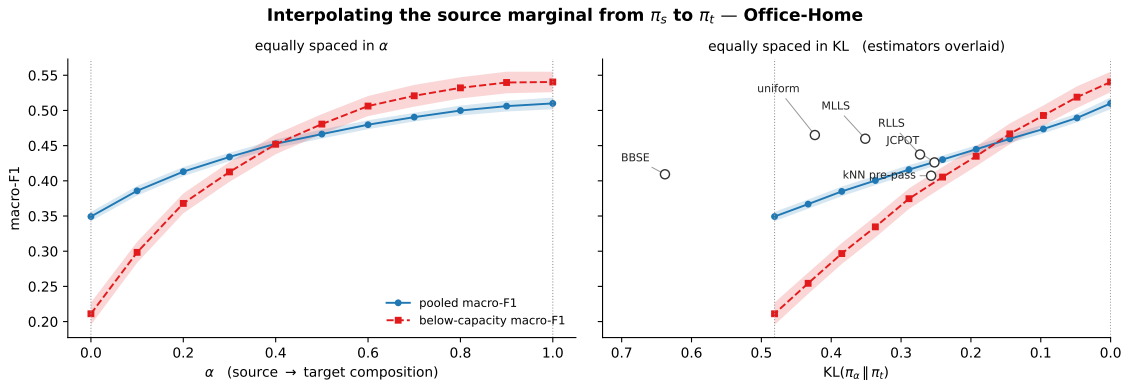


Figure 2: **Office-Home.** Macro-F1 along the interpolation path from the source composition ($\alpha = 0$) to the target composition ($\alpha = 1$). The right panel plots the same results against KL divergence to the target; circles mark the composition estimates evaluated in Sec. 4.3.

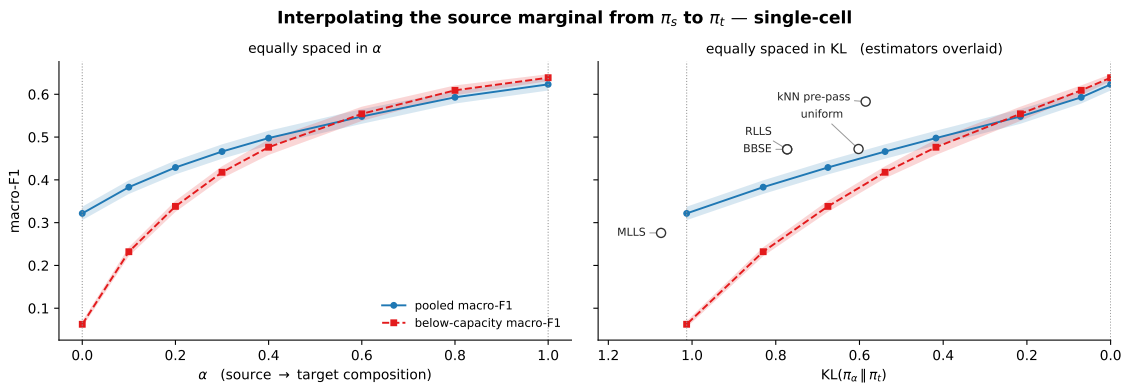


Figure 3: **Single-cell.** Macro-F1 along the interpolation path from the source composition ($\alpha = 0$) to the target composition ($\alpha = 1$). The right panel plots the same results against KL divergence to the target; circles mark the composition estimates evaluated in Sec. 4.3.

5 Discussion and Conclusion

We asked whether the poor transfer of rare classes under composition shift can arise from the OT formulation itself, rather than only from the representation. Our results show that it can. The source marginal imposes a class-level transport capacity, so a class can be starved when its target demand exceeds its available mass, even when the embedding separates it well. The displaced mass is absorbed by classes with greater capacity rather than by geometrically similar classes, and the resulting errors can remain highly confident. Correcting the marginal removes this coverage-dependent component of the error, while the residual tracks class separability. Thus, marginal capacity is a distinct source of transfer error, not simply another manifestation of poor representation quality.

The location of this failure also explains why downstream corrections are limited. Once the coupling has allocated too little mass to a class, changing the read-out, smoothing the plan, or relaxing the marginal constraint cannot reliably recover the missing mass. The effective intervention is instead to change the class-level mass allocation itself. This also clarifies the practical implication of the correction: the target composition, rather than a generic preference for rare classes, should determine the transport marginal.

Limitations and Future Work The composition-matched marginal uses the true target composition and therefore provides an upper bound; estimated compositions weaken the capacity–geometry separation. Future work could develop more accurate target-composition estimators to improve the practical correction. Our analysis treats the representation as fixed; extending OT label transfer to jointly learned, multimodal representations could provide a complementary way to improve cross-domain separability. Finally, the controlled Office-Home experiments provide the clearest evidence for separating coverage from separability, while the single-cell datasets contain fewer below-capacity classes; broader domain shifts and coupling-level diagnostics could test and detect the same failure more generally.

Conclusion OT label transfer under composition shift is limited by two distinct factors: the representation determines which classes can be distinguished, while the transport marginal determines how much mass each class can receive. A class can therefore be under-predicted because the coupling cannot allocate enough mass even when the representation is informative, producing confident errors that downstream read-outs cannot recover. Matching the marginal to the target composition removes this capacity-driven component and reveals the residual geometric error, but the correction is useful primarily where coverage is limiting and requires an estimate of the target composition.

REFERENCES

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22:1657–1661, 2025. doi: 10.1038/s41592-025-02772-6.

Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *PMLR*, pp. 222–232, 2020.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In *International Conference on Learning Representations (ICLR)*, 2019.

Ihab Bendidi, Shawn Whitfield, Kian Kenyon-Dean, Hanene Ben Yedder, Yassir El Mesbahi, Emmanuel Noutahi, and Alisandra Kaye Denton. Benchmarking transcriptomics foundation models for perturbation analysis: One PCA still rules them all. *arXiv preprint arXiv:2410.13956*, 2024. NeurIPS 2024 AIDrugX Workshop.

Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R Vanderburg, Åsa Segerstolpe, Meng Zhang, et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nature methods*, 18(11):1352–1362, 2021.

Rebecca Boiarsky, Nalini Singh, Alejandro Buendia, Gad Getz, and David Sontag. A deep dive into single-cell RNA sequencing foundation models. *bioRxiv*, 2023. doi: 10.1101/2023.10.19.563100.

Maria Brbić, Kaidi Cao, John W. Hickey, Yuqi Tan, Michael P. Snyder, Garry P. Nolan, and Jure Leskovec. Annotation of spatially resolved single-cell data with STELLAR. *Nature Methods*, 19(11):1411–1418, 2022. doi: 10.1038/s41592-022-01651-8.

Charlotte Bunne, Geoffrey Schiebinger, Andreas Krause, Aviv Regev, and Marco Cuturi. Optimal transport for single-cell and spatial omics. *Nature Reviews Methods Primers*, 4:58, 2024. doi: 10.1038/s43586-024-00334-2.

Kai Cao, Qiyu Gong, Yiguang Hong, and Lin Wan. A unified computational framework for single-cell data integration with optimal transport. *Nature Communications*, 13:7419, 2022. doi: 10.1038/s41467-022-35094-8.

Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40:1458–1466, 2022. doi: 10.1038/s41587-022-01284-4.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

Lenaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018. doi: 10.1090/mcom/3303.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865, 2017. doi: 10.1109/TPAMI.2016.2615921.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: Toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21: 1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

Damien Dablain, Bartosz Krawczyk, and Nitesh V Chawla. Deepsmote: Fusing deep learning and smote for imbalanced data. *IEEE transactions on neural networks and learning systems*, 34(9):6390–6404, 2022.

C. Domínguez Conde, C. Xu, L. B. Jarvis, D. B. Rainbow, S. B. Wells, T. Gomes, others, and S. A. Teichmann. Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science*, 376(6594): eabl5197, 2022. doi: 10.1126/science.abl5197.

Mourad El Hamri, Younès Bennani, and Issam Falih. Label propagation through optimal transport. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021.

Kilian Fatras, Thibault Séjourné, Nicolas Courty, and Rémi Flamary. Unbalanced minibatch optimal transport; applications to domain adaptation. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 3186–3197, 2021.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, pp. 1321–1330, 2017.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21:1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

Doron Haviv, Ján Remšík, Mohamed Gatie, Catherine Snopkowski, Meril Takizawa, Nathan Pereira, John Bashkin, Stevan Jovanovich, Tal Nawy, Ronan Chaligne, et al. The covariance environment defines cellular niches for spatial inference. *Nature Biotechnology*, 43(2):269–280, 2025.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

Amanda Janesick, Robert Shelansky, Andrew D. Gottscho, Florian Wagner, Stephen R. Williams, Morgane Rouault, et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications*, 14:8353, 2023. doi: 10.1038/s41467-023-43458-x.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Kasia Z. Kedzierska, Lorin Crawford, Ava P. Amini, and Alex X. Lu. Zero-shot evaluation reveals limitations of single-cell foundation models. *Genome Biology*, 26:101, 2025. doi: 10.1186/s13059-025-03574-x.

Dominik Klein, Giovanni Palla, Marius Lange, Michal Klein, Zoe Piran, Manuel Gander, Laetitia Meng-Papaxanthos, Michael Sterr, Lama Saber, Changying Jing, et al. Mapping cells through time and space with moscot. *Nature*, 638(8052):1065–1075, 2025.

Bin Li, Wen Zhang, Chuang Guo, Hao Xu, Longfei Li, Minghao Fang, Yinlei Hu, Xinye Zhang, Xinfeng Yao, Meifang Tang, et al. Benchmarking spatial and single-cell transcriptomics integration methods for transcript distribution prediction and cell type deconvolution. *Nature methods*, 19(6):662–670, 2022.

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pp. 3122–3130, 2018.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, others, and Fabian J. Theis. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40:121–130, 2022. doi: 10.1038/s41587-021-01001-7.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 427–436, 2015.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

Ievgen Redko, Nicolas Courty, Rémi Flamary, and Devis Tuia. Optimal transport for multi-source domain adaptation under target shift. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89 of *PMLR*, pp. 849–858, 2019.

Yanay Rosen, Yusuf Roohani, Ayush Agrawal, Leon Samotorcan, Tabula Sapiens Consortium, Stephen R. Quake, and Jure Leskovec. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, 2023. doi: 10.1101/2023.11.28.568918.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002. doi: 10.1162/089976602753284446.

Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. *arXiv preprint arXiv:1910.12958*, 2019.

Thibault Séjourné, Gabriel Peyré, and François-Xavier Vialard. Unbalanced optimal transport, from theory to numerics. *Handbook of Numerical Analysis*, 24:407–471, 2023.

Bharath Srinivasan. A guide to the michaelis–menten equation: Steady state and beyond. *The FEBS Journal*, 289(20):6086–6098, 2022. doi: 10.1111/febs.16124.

Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 177(7):1888–1902, 2019. doi: 10.1016/j.cell.2019.05.031.

Remi Tachet des Combes, Han Zhao, Yu-Xiang Wang, and Geoffrey J. Gordon. Domain adaptation with conditional distribution matching and generalized label shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Alejandro Tejada-Lapuerta, Anna C Schaar, Robert Gutgesell, Giovanni Palla, Lennard Halle, Mariia Minaeva, Larsen Vornholz, Leander Dony, Francesca Drummer, Till Richter, et al. Nicheformer: a foundation model for single-cell and spatial omics. *Nature methods*, 22(12):2525–2538, 2025.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, others, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, 2023. doi: 10.1038/s41586-023-06139-9.

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5018–5027, 2017.

Tom Viering and Marco Loog. The shape of learning curves: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7799–7819, 2023. doi: 10.1109/TPAMI.2022.3220744.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. CellPLM: Pre-training of cell language model beyond single cells. In *International Conference on Learning Representations (ICLR)*, 2024.

Aaron Wenteler, Martina Occhetta, Nikhil Branson, Magdalena Huebner, Victor Curean, WT Dee, WT Connell, Alex Hawkins-Hooker, Siu Pui Chung, Yasha Ektefaie, et al. Perteval-scfm: benchmarking single-cell foundation models for perturbation effect prediction. *BioRxiv*, pp. 2024–10, 2024.

Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I. Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular Systems Biology*, 17(1):e9620, 2021. doi: 10.15252/msb.20209620.

REPRODUCIBILITY STATEMENT

Apps. B through C give the derivation with its assumptions and scope as well as the datasets, folds, budgets, solver settings and metric definitions. The analysis in App. F re-reads logged per-class recalls from the runs behind Table 4 and solves no new transport problem. Janesick, SpatialCorpus and Office-Home are public. The code can only be shared upon the acceptance of the paper due to the internal policies of our organization.

A EXTENDED RELATED WORK

Transport-based label transfer. Unpaired modalities are aligned with coupled autoencoders and OT (Cao et al., 2022), graph-linked embeddings (Cao & Gao, 2022), anchor transfer (Stuart et al., 2019), semi-supervised generative models (Xu et al., 2021), geometric deep learning (Brbić et al., 2022) and niche-covariance models (Haviv et al., 2025); reference mapping transfers atlas labels (Lotfollahi et al., 2022; Domínguez Conde et al., 2022); foundation models supply reusable embeddings (Tejada-Lapuerta et al., 2025; Cui et al., 2024; Hao et al., 2024; Theodoris et al., 2023; Rosen et al., 2023; Wen et al., 2024). None characterizes, per class, when transfer is feasible, and none identifies the source marginal as the controlling object.

Entropic and unbalanced OT. Entropic OT (Cuturi, 2013; Peyré & Cuturi, 2019) and its unbalanced relaxations (Chizat et al., 2018; Séjourné et al., 2019; 2023; Fatras et al., 2021) underpin single-cell mapping (Biancalani et al., 2021; Klein et al., 2025; Bunne et al., 2024) and domain adaptation (Courty et al., 2017). The unbalanced relaxation is motivated throughout as the response to composition mismatch. Sec. 3.3 measures that response against specifying the marginal correctly and finds the latter roughly twice as effective on the classes that have a deficit.

Label shift. See Sec. 4.3 for positioning. Generalized label shift (Tachet des Combes et al., 2020) relaxes the invariance assumption that Sec. 4.3 finds violated under a measurement-process gap. Abundance-aware spot deconvolution (Li et al., 2022) is the nearest acknowledgement in spatial transcriptomics that composition is first-class, but it estimates mixture proportions per location rather than labels per point.

Simple baselines, confident errors, saturation laws. Foundation-model embeddings frequently fail to beat simple baselines (Kedzierska et al., 2025; Boiarsky et al., 2023; Bendidi et al., 2024; Ahlmann-Eltze et al., 2025; Wenteler et al., 2024); Sec. 4 shows one such reported gap dissolving under matched corrections. Classifiers are miscalibrated off-distribution (Guo et al., 2017; Nguyen et al., 2015; Ovadia et al., 2019), and selective prediction (Geifman & El-Yaniv, 2017), softmax and OOD baselines (Hendrycks & Gimpel, 2017) and conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021) act at the output, where Sec. 3.2 finds these errors invisible. Equation 3 is a rectangular hyperbola of Michaelis and Menten form (Srinivasan, 2022), derived from the marginal rather than fitted as an empirical scaling law (Hestness et al., 2017; Kaplan et al., 2020; Viering & Loog, 2023). Optimal-Transport Propagation, which we test in App. E, diffuses labels over a graph built from the coupling and is the strongest plan-side read-out we are aware of.

B THE SATURATING LAW AND ITS SCOPE

For a target point x of true class t, the entropic row of the optimal plan is $T_{x\cdot} \propto \exp(-(C_{x\cdot} - f_x - g)/\varepsilon)$. Aggregating by source label, $\sum_{j\in t} T_{xj} \propto n_t^{\mathrm{s}}\, \bar{e}_t$ with $\bar{e}_t \approx e^{-\sigma^2/\varepsilon}$ for the true class and $\bar{e}_c \approx e^{-d^2/\varepsilon}$ for the nearest well-covered class c, where σ is within-class scatter and d centroid separation. A soft read-out

assigns t with probability $n_t^{\mathrm{s}}\bar{e}_t/(n_t^{\mathrm{s}}\bar{e}_t + n_c^{\mathrm{s}}\bar{e}_c)$; dividing by n_t^{t} gives

$$\mathrm{recall}(t) \;\approx\; \frac{a\, r_t}{r^\star + r_t}, \qquad r^\star = \kappa\, e^{-(d^2-\sigma^2)/\varepsilon}. \tag{3}$$

Note the two factors: a capacity term in r_t and a geometry term in $d^2 - \sigma^2$. The double dissociation of Sec. 4.2 is a test of exactly this structure, and it is the part of Eq. 3 that survives outside the single-cell setting.

Assumptions and scope. The derivation is written for a *soft* read-out while the deployed one is an $\arg\max$, and in the $\tau \to \infty$ limit, absorbing the marginal penalty into the potentials; with finite τ mass conservation is soft, so Eq. 3 is the hard-constraint statement. The τ sweep is consistent with this, since the pathology is strongest exactly at $\tau \to \infty$. The form does not transfer to Office-Home, where the median R^2 is 0.276 for transport and 0.142 for the vote, $r^\star$ is undefined for roughly half the classes under transport, the transport-to-vote ratio is $1.26\times$, and no class sits at chance (Table 5). We therefore present it as a single-cell instantiation rather than a general law. We do not quote a single-cell threshold ratio: the logged Janesick fits predate the $a \le 1$ convention, no post-refit value exists in a logged run, and at the default τ the fits are not identified at all (Sec. 3.3, App. E). The single-cell claim therefore rests on the fit-free resolvability count of Sec. 3.1, which is computed before any fit is attempted. What fails on Office-Home is the parametric *form*. The structure it is derived from, a capacity factor in r times a geometry factor, is tested directly in Sec. 4.2 and holds on both domains. The qualitative gap has the same explanation as the chance-level classes of Sec. 3.1: a 65-class tail has no dominant absorber, so no class is starved hard enough to sit at chance.

Table 5: The saturating law fitted on Office-Home, transport read-out against a kNN vote. Same protocol as the single-cell fits: per class, $\mathrm{recall}(t) \approx a\, r_t/(r^\star + r_t)$, ceiling constrained to $a \le 1$, and any fit landing within 10^{-3} of that bound declined. No fit is pinned here (the largest ceiling is 0.9951), so nothing is declined for that reason. The form does not transfer: the fits are poor, a threshold is undefined for roughly half the classes under transport, the transport-to-vote ratio is close to unity with a range spanning both sides of it, and no class sits at chance. Single-cell fits are not tabulated alongside, because the logged Janesick fits predate the $a \le 1$ convention; Sec. 3.1 rests the single-cell claim on the fit-free resolvability count instead, which no fitting convention can affect.

	transport	kNN vote
classes resolvable	$65/65$	$64/65$
classes with a finite $r^\star$	$36/65$	$47/65$
median R^2	0.276	0.142
median $r^\star$	0.142	0.108
median ceiling a	0.657	0.606
fits declined at the $a \le 1$ bound	$0/36$	$0/47$
classes at chance at every coverage tested	$0/65$	$0/65$
paired over the 34 classes resolving under both		
median $r^\star$ ratio, transport / vote	$1.263\times$	
range of the ratio	$[0.225\times,\ 3.560\times]$	

C Data, protocol and metrics

Jint (in-house). Twelve paired spatial and dissociated slides; $802{,}132$ dissociated points $\times$ $18{,}051$ features and $27{,}238$ to $273{,}070$ points per spatial slide $\times$ $5{,}001$ features; seven reconciled classes; annotations on both sides with zero missing labels. Main runs use folds carrying a below-capacity class, an uncapped $195{,}598$-point source, three seeds, and a converged Sinkhorn–Knopp and majorization–minimization chain,

the two agreeing in objective to eight decimals at 5×10^5 iterations, with every run recording its converged backend. The below-capacity stratum is three (fold, class) cells.

Janesick (public, GSE243275). Pooled $3'$ and $5'$ dissociated source (22,969 and 12,388 points $\times$ 36,601 features) to a 313-feature spatial target (167,780 and 118,752 points, two replicates); published annotations on all four files. Per-class law fits use 48 observations; the coverage gradient spans three orders of magnitude and includes an above-capacity control class ($r = 0.601$).

SpatialCorpus (public atlas). Five tissues, two species; source and target are partitions of one dissociated file, so compositions agree by construction. Used as the composition-shift negative control.

Office-Home. Four domains, 65 classes, 15,588 images; CLIP ViT-B/32 features, with ResNet-50 followed by PCA as a second backbone. Target composition long-tailed at 100:1; per-class r drawn log-uniform on $[0.01, 2]$ and n_t^{s} derived from the draw, so the realised ratio matches the design to 2×10^{-16}, where dividing two independent long tails instead gives $[0.014, 39]$. Achieved $r \in [0.020, 5.50]$ with $235/780$ class-observations below capacity, and $\rho(\log r, \text{separability}) = +0.108$. Four domain pairs, three seeds. Each Office-Home experiment draws its own composition profile, held fixed across all arms within that experiment, so every comparison we report is a within-experiment contrast and we do not compare arms across experiments. ε was recalibrated on CLIP features rather than carried over.

Source budgets. Training-free read-outs use the uncapped pool. Trained methods cannot fit it and use a capped proportional source; composition is preserved while absolute counts scale, so the two are reported separately and never ranked together. Backbone-invariance runs use a 40,000-point source because the PCA basis needs counts, and they test reproduction on another representation rather than an absolute $r^\star$.

Target-composition design, by result. Natural composition is used for the marginal and estimator panels (Table 4), the matched-correction comparison (Table 15), every Office-Home result, and the mechanism counts and capacity partials (Table 2). A fixed-per-class target is used, and enters only as a within-run contrast, for the four-solver comparison and the detector suite (Table 13), the backbone-invariance run, the ε and τ sweeps and the permutation control on single-cell data (Sec. 3.3), the read-out and detection experiments of App. E, the ratio-versus-count pair, the minority-synthesis nulls, and the Janesick per-class curve fits, where the coverage cap is part of the sweep design. In each of the latter the target is held fixed across the arms compared, so the ordering is unaffected, while absolute levels are not comparable across the two designs.

Metrics. Macro recall is the unweighted mean of per-class recall, macro-F1 the unweighted mean of per-class F1, and micro recall overall per-point accuracy. Absorption is the fraction of a below-capacity class's target mass received by a single other class. Detector selectivity is the ratio of below-capacity error rate among flagged points to that among unflagged points, and coverage is the fraction of points not abstained on. Domain identity is close to linearly separable in the embedding, so we remove it by per-domain mean centering before computing C; this is standard and is itself worth a large fraction of macro recall (0.320 to 0.426 from raw to centered, absorption 0.994 to 0.974). Every arm in every comparison shares it.

D THE CORRECTION, SEEN IN THE PLAN

The results in the body are aggregate statistics. This appendix shows the same effect on a single fold, in the coupling itself and in the embedding, with the source marginal as the only thing that differs between arms. Both figures use Jint fold G1, seed 42, a 195,893-point source and a 5,000-point target at natural composition, with identical ε and $\tau = 1$ throughout. Endothelial ($r = 0.141$) is the one class below capacity

on this fold; Stromal ($r = 0.32$) improves by a similar amount but is not below capacity, and we do not describe it as such.

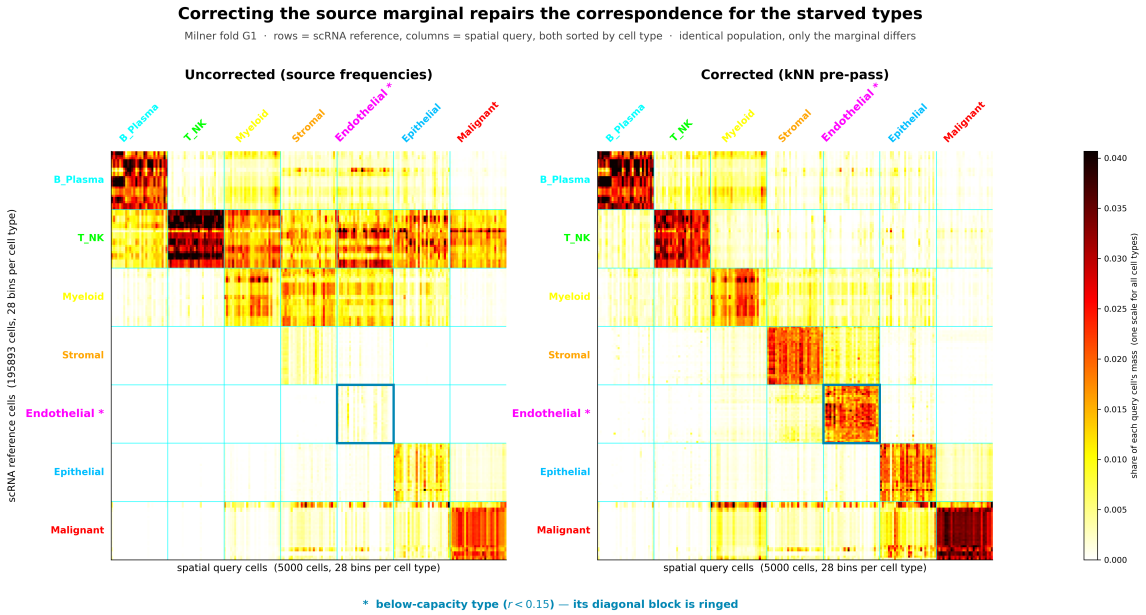


Figure 4: **The failure is in the coupling, and the marginal repairs it.** Entropic unbalanced OT couplings between the source (rows) and the target (columns), both sorted by class. Each class is allocated the same number of bins on each axis, so a class holding 0.3% of the source is as legible as one holding half of it, and colour is the share of each target point's transported mass falling in a given source bin, on one scale shared by both panels and all classes. *Left:* the source marginal is the source's own class frequencies. *Right:* the marginal is rebuilt as $\mu_j = \hat{\pi}_{y_j}/n^{\mathrm{s}}_{y_j}$ from a kNN pre-pass estimate of the target composition, which uses no target labels. Everything else is identical. The horizontal band on the T_NK rows in the left panel is the absorption mechanism of Sec. 3.2 made visible: target points of a starved class are transported to a class that has capacity. Endothelial, the one class below capacity here, is starred and its diagonal block is ringed; uncorrected it sends 46.9% of its mass to T_NK and 3.9% to itself, and corrected it sends 51.3% to itself. Aggregated over the whole target, the T_NK source rows absorb 47.8% of all transported mass before correction and 14.4% after.

Figure 5 shows the same fold in the embedding rather than the plan.

Per-class detail, and what the correction costs. Table 6 gives recall and diagonal coupling mass for every class on this fold. Two things are worth stating plainly. The gain is largest on the classes with the least coverage and shrinks monotonically as r rises, which is the dose-response of Sec. 4.1 on a single fold. And the correction is not free: T_NK, the class doing the absorbing, falls from 0.948 to 0.768, since the marginal redistributes mass rather than creating it. That is why the adjudicating rule is macro-F1 with a micro floor rather than below-capacity recall alone.

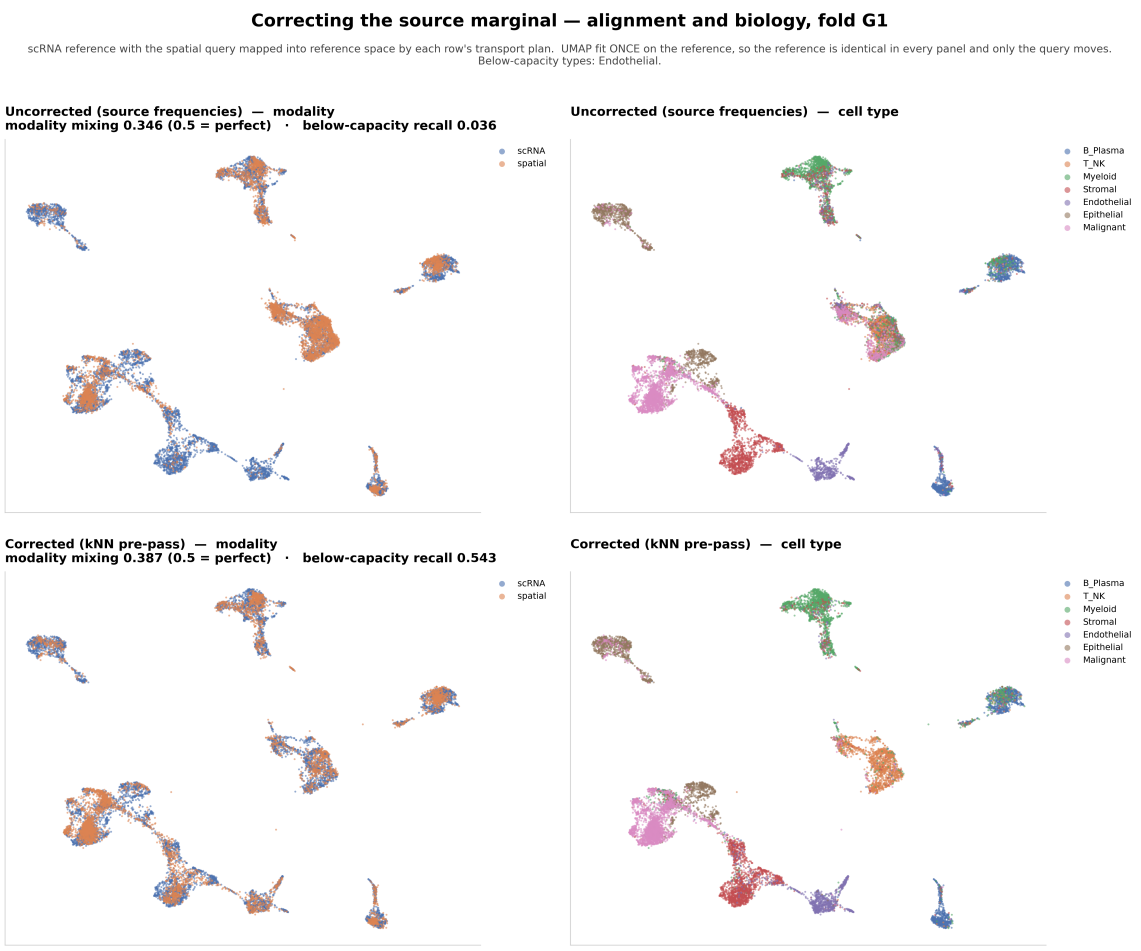


Figure 5: **The correction acts on class correspondence, not on domain alignment.** UMAP of the source together with the target mapped into source space by each row's transport plan through barycentric projection. The reducer is fit once, on the source alone, and both rows are transformed into that fixed embedding, so the source points are identical in every panel and only the target moves; any difference between rows is therefore attributable to the plan and not to a refit. *Top row:* source-frequency marginal. *Bottom row:* composition-matched marginal from the kNN pre-pass. *Left column:* points coloured by domain, which measures alignment. *Right column:* points coloured by class, which checks that the biology is preserved. Domain mixing, the fraction of each point's fifteen nearest neighbours drawn from the other domain on the plotted coordinates, where 0.5 is perfect interleaving, improves only from 0.346 to 0.387: the two domains were already reasonably interleaved, and alignment is not what the marginal repairs. The class panels show what does change, with the Endothelial population diffuse and largely mislabelled before correction and consolidated after, and below-capacity recall rising from 0.036 to 0.543.

Table 6: Jint fold G1, per class, ordered by coverage ratio. *Diagonal mass* is the share of a class's target mass landing on its own source block, read off the plotted matrices of Fig. 4. Endothelial is the only class below capacity on this fold.

class	r	recall before	recall after	diagonal mass
Endothelial (below capacity)	$\mathbf{0.141}$	0.036	$\mathbf{0.543}$	$3.9\% \rightarrow 51.3\%$
Stromal	0.320	0.094	0.619	$8.7\% \rightarrow 55.8\%$
Epithelial	1.903	0.238	0.514	$23.7\% \rightarrow 48.6\%$
Malignant	2.305	0.561	0.896	$53.1\% \rightarrow 84.9\%$
Myeloid	3.135	0.314	0.442	$32.5\% \rightarrow 43.3\%$
B_Plasma	5.943	0.711	0.853	$69.9\% \rightarrow 83.9\%$
T_NK	19.875	0.948	$\mathbf{0.768}$	$93.5\% \rightarrow 73.4\%$
macro / micro		0.415 / 0.443	0.662 / 0.690	
domain mixing		0.346	0.387	

E READ-OUT, MARGINAL PENALTY AND ENTROPIC REGULARIZER, IN FULL

This appendix reports in full the three tunings summarized in Secs. 3.1 to 3.3. Each was specified before it was run, and each has the same shape: the knob moves the deficit part of the way and none closes it.

Read-outs of a fixed coupling. One converged plan per unit, with every read-out applied to that identical plan, so no two arms can differ by a solver iteration. The kNN vote is the deliberate exception, since it ignores the plan entirely and is the no-coupling reference. Table 7 gives below-capacity recall for all eleven arms on three datasets.

Table 7: Read-out battery, below-capacity recall under the natural marginal. Every arm except kNN consumes the same converged plan. Hard, tempered and soft read-outs are identical to four decimals; propagation buys a little and twenty steps buy nothing over one; the only plan read-outs that help are the two that divide by source class count, which is the marginal correction relocated after the solve. $n = 12$ (Jint), 2,412 (rep1), 1,926 (rep2).

family	read-out	Jint	Janesick rep1	Janesick rep2
hard	argmax	0.040	0.071	0.211
tempered	softmax T=1	0.040	0.071	0.211
tempered	softmax T=0.05	0.040	0.071	0.211
soft	soft posterior	0.040	0.071	0.211
propagating	OTP, 1 step	0.062	0.085	0.254
propagating	OTP, 20 steps	0.063	0.086	0.259
propagating	OTP no gate, 1 step	0.062	0.085	0.254
propagating	OTP no gate, 20 steps	0.063	0.086	0.259
prior-corrected	argmax $/\, n_t^{\mathrm{s}}$	$\mathbf{0.199}$	$\mathbf{0.180}$	0.383
prior-corrected	$\mathrm{OTP}_{20} \,/\, n_t^{\mathrm{s}}$	0.188	0.190	$\mathbf{0.394}$
plan-free	kNN vote	$\mathbf{0.395}$	0.165	$\mathbf{0.465}$

Two conclusions follow, and a third that we pre-registered and must withdraw. Softening changes nothing: a soft posterior over a plan whose mass sits on the wrong classes is a soft posterior over the wrong classes. And a read-out that never touches the plan beats every read-out that does, on two of three datasets. The withdrawn prediction concerns the entropy gate inside OTP, which we expected to suppress rare classes,

since confident hallucination implies it should clamp confident but wrong points early. Gated and ungated variants agree to four decimals everywhere in Table 7, so the gate contributes nothing and we report it as a null.

The floor is in the coupling, not in its interpretation. Table 8 settles the question without appeal to any baseline. The share of transported mass reaching a below-capacity class is under two percent on every dataset, and the plain argmax already converts almost all of it into correct labels, so there is no headroom for a cleverer read-out to exploit. The prior-corrected arms exceed that ceiling only because they do not read the plan proportionally: they rescale by source class size, which is Eq. 2 applied after the solve rather than before.

Table 8: The coupling-mass ceiling, natural marginal. *Coupling mass* is the share of total transported mass landing on a below-capacity class and *expected recall* is what a perfect read-out of that plan could achieve; both are read-out-independent properties of the coupling. The last column is what argmax actually extracts as a fraction of that ceiling.

dataset	coupling mass	expected recall	argmax recall	extracted
Jint	0.0041	0.041	0.040	96%
Janesick rep1	0.0190	0.077	0.071	93%
Janesick rep2	0.0160	0.183	0.211	115%

Which axis moves the number, and a second collapse. Holding argmax fixed while the marginal changes lifts below-capacity recall by $15.4\times$, $3.7\times$ and $2.8\times$ on Jint and the two replicates, against $5.0\times$, $2.7\times$ and $1.9\times$ for the best read-out at a fixed marginal; and the read-out column is generous, since its best entries are the two prior-corrected arms. Restricted to read-outs that read the plan faithfully the lift is 1.2 to $1.6\times$. Under the corrected marginal the read-outs also converge, from a $5.0\times$ spread on Jint to $1.2\times$, which is the same collapse the marginal–τ interaction shows on the τ axis: both knobs have leverage only while the marginal is wrong. Prior correction is not free, either. On replicate 1 it costs above-capacity recall $0.309 \to 0.230$ while gaining $0.071 \to 0.180$ below capacity, which is why the adjudicating rule is macro-F1 with a micro floor rather than below-capacity recall alone.

Fitted thresholds for this battery. On replicate 1 the median transport-to-vote threshold ratio is $18.8\times$ for argmax and the identical soft arms, $17.3\times$ for one-step OTP and $15.8\times$ for twenty steps, against $0.13\times$ and $0.05\times$ for the prior-corrected arms and $1.00\times$ for the vote by construction. Replicate 2 yields *no* identified fit for any plain read-out. These ratios rest on one to three identified fits each, with most (slice, class) fits declined because the curve sits at chance across the entire swept range. We therefore quote them as direction rather than magnitude; the refusals are the more honest headline, since a read-out that never rises above chance has no threshold to fit.

Marginal penalty. Table 9 sweeps τ across five orders of magnitude on both datasets. The ordering is monotone on each independently, the balanced limit is the worst arm on both, and absorption tracks the starved classes exactly. Fitting Eq. 3 per arm on Janesick also shows that the threshold is identifiable only once τ is relaxed, which is one reason App. B reports no single-cell threshold ratio. Asymmetric penalties on the two marginals were not run; we predict them at or below the balanced-marginal point, since symmetric relaxation already saturates at about half of the shape-correction effect.

Marginal $\times$ τ, the interaction. Table 10 crosses three marginals with five τ on both domains. The design was specified before it was run, with three pre-registered questions. The first asked whether the τ optimum

Table 9: The τ sweep on both datasets. *Identified* $r^\star$ is the number of the fourteen Janesick classes yielding a fit that is not declined; at the default τ and tighter, every fit is declined with its ceiling pinned and the note that the curve does not saturate within the observed window.

	Janesick	Jint			Janesick
τ	macro	macro	absorption	Endothelial / Stromal	identified $r^\star$
0.01 (most relaxed)	**0.451**	**0.586**	0.656	0.217 / 0.435	6/14
0.1	0.399	0.519	0.863	0.091 / 0.222	4/14
1 (default)	0.306	0.426	0.974	0.020 / 0.058	0/14
10	0.267	0.407	0.985	0.012 / 0.039	0/14
∞ (balanced)	0.262	0.404	**0.987**	0.012 / 0.036	0/14

moves toward the enforced limit once the marginal is correct. It does on the single-cell oracle row, monotonically across all five values, and it does not on Office-Home, whose oracle row is flat; since neither row's span is significant, we do not claim it. The leverage collapse is the sound form of the same prediction and it holds decisively on both datasets. The second asked whether an imperfect estimate wants some relaxation as a hedge against estimation error. It does not. The Office-Home deployable row peaks at the most relaxed τ, exactly like the default row, and the single-cell deployable row peaks at $\tau = 10$ by 4.5×10^{-5} against a standard error of 0.072, which is an argmax tie in a flat row rather than a sweet spot. There is no middle-ground regime on either domain. The third asked whether the combination beats the better single knob, and it does not, by $+0.0007$ and $+0.0100$ at the oracle and $+0.0270$ and $+0.0025$ at the deployable estimator.

Bracketing the relaxed end. Two rows in Table 10 peak at $\tau = 0.01$, the edge of the grid, so we extended both domains by one decade. No τ=0.001 minus τ=0.01 difference is significant on either domain, the largest reaching $z = +0.81$, and the Office-Home deployable row now has an interior optimum at $\tau = 0.01$. Rows that still peak at the new edge do so by amounts indistinguishable from zero, so a further decade of relaxation cannot change any conclusion. We report leverage on the pre-registered five-τ grid and treat the extension as a bracketing check.

Absorption gives the mechanism. On single-cell data at the default marginal with the constraint enforced, 98.0% of below-capacity points are absorbed by a covered class; the oracle marginal at the same τ cuts that to 28.4%. Within the default row relaxation is what suppresses absorption, from 0.98 to 0.69, whereas within the oracle row enforcement is, from 0.50 to 0.28. The direction in which τ moves absorption reverses with the correctness of the marginal, which is why a τ sweep run on a wrong marginal was always going to look as though rigidity were the problem.

Entropic regularizer and detection. Seven values of ε, three folds and three seeds, on a fixed-per-class target matching the main detector run exactly, so selectivities sit on the same scale. Table 11 shows the Pareto block: selectivity rises by a factor of seven across the sweep while below-capacity recall collapses to zero, so the detector improves precisely as there stops being anything to detect. The confidence columns give the reason. Softening lowers confidence on correct and on incorrect points together, and expected calibration error tracks the latter exactly, so the widening gap is a rescaling rather than a separation. The sweep covers every scorable unit: one fold carries no below-capacity class and is skipped, so the denominator is smaller than the fold count elsewhere in this paper, and the selectivity at the default ε is therefore not directly comparable to the main detector suite.

Table 10: Marginal $\times$ τ, below-capacity macro-F1. *Deployable* is JCPOT on Office-Home and the kNN pre-pass on single-cell data, each the leading estimator in its own panel; *oracle* uses the true composition and is an upper bound rather than a method. *Span* is best minus worst within the row, with its z. On a wrong marginal τ is a large lever; on a corrected one no setting is distinguishable from any other. Office-Home is 4 domain pairs $\times$ 3 seeds and single-cell is 3 folds $\times$ 3 seeds, both complete. A further arm at $\tau = 0.001$ leaves every conclusion unchanged (App. E).

domain	marginal	τ=0.01	0.1	1	10	∞	span	z
Office-Home	default	**0.450**	0.359	0.217	0.193	0.190	**0.260**	11.8
	deployable	**0.477**	0.451	0.431	0.426	0.426	0.051	1.8
	oracle	0.528	**0.542**	0.541	0.538	0.537	0.014	0.5
single-cell	default	**0.408**	0.203	0.057	0.044	0.042	**0.366**	8.7
	deployable	0.547	0.542	0.552	**0.555**	0.555	0.013	0.2
	oracle	0.549	0.572	0.634	0.643	**0.644**	0.096	1.3

Table 11: The ε sweep, as a multiple of the calibrated default. No setting has both meaningful below-capacity recall and a detector clearing the pre-registered bar. The two smallest values produce a numerically degenerate plan. *Sel.* is the selectivity of the best output-side detector; *conf.* are mean confidences on correct covered points and on wrong below-capacity points. Six scorable units: one fold has no below-capacity class, so the detector suite is undefined there and it contributes no rows. The bar is cleared on at most three of six units at any ε, never all six.

ε	degenerate	below-cap. recall	macro	sel.	conf. correct	conf. wrong
0.01$\times$	yes	0.000	0.143	1.00	—	—
0.03$\times$	yes	**0.132**	0.312	1.15	—	—
0.1$\times$	no	0.027	0.441	2.94	0.974	0.874
0.3$\times$	no	0.027	**0.442**	2.89	—	—
1$\times$	no	0.024	0.438	3.28	0.927	0.744
3$\times$	no	0.002	0.408	5.15	0.817	0.554
10$\times$	no	0.000	0.278	**23.33**	0.632	0.416

F THE DECOMPOSITION UNDER A DEPLOYABLE CLASS PRIOR

Sec. 4.2 defines removable and residual against the oracle marginal. This appendix recomputes both with each deployable estimator substituted for the oracle, on the same 780 Office-Home class-observations and the same rank-normalisation. No optimal transport problem was re-solved; the analysis reads the logged per-class recalls of arms that already appear in Table 4. The achieved $\rho(\log r, \text{separability}) = +0.108$ is a property of the design and is therefore identical across arms.

Reading the table. Two of the four relations are stable. Removable is governed by coverage on every arm, at raw ρ from -0.487 to -0.607 with every arm at $p < 10^{-47}$, and residual is governed by separability on every arm, at raw ρ from -0.330 to -0.519 with every arm at $p < 3{\times}10^{-21}$. The two near-zero cells at the oracle behave differently from one another. The residual's coverage cell becomes significantly negative under every estimate, which is the leak discussed in Sec. 4.2. The removable's separability cell becomes significantly positive under every estimate, from $+0.108$ to $+0.238$ raw, which is why we reserve the word dissociation for the oracle. We also note that the oracle's own partial for this cell is $+0.135$ rather than the raw -0.009, because removable is strongly driven by $\log r$ and the two predictors retain a small positive association by construction.

Table 12: The decomposition under seven priors, Office-Home, $n = 780$. Cells are partial Spearman ρ with the other predictor controlled. Removable stays coverage-governed and residual stays separability-governed on every arm, while the near-zero cells at the oracle become non-zero under every estimate.

prior for π	removable $\sim \log r$	removable $\sim$ sep	residual $\sim$ sep	residual $\sim \log r$
oracle (published)	$\mathbf{-0.813}$	$+0.135$	$\mathbf{-0.338}$	$+0.034$
uniform	-0.573	$+0.316$	-0.502	-0.157
kNN pre-pass	-0.517	$+0.263$	-0.374	$\mathbf{-0.435}$
JCPOT	-0.516	$+0.188$	-0.313	-0.287
BBSE	-0.541	$+0.282$	-0.422	-0.296
RLLS	-0.538	$+0.289$	-0.429	-0.305
MLLS	$\mathbf{-0.655}$	$+0.385$	-0.513	$\mathbf{-0.072}$

Uniform balancing is not an estimate. Its leak of -0.157 is smaller than that of several genuine estimators, but uniform is a fixed balanced prior rather than an estimate of π, and the Office-Home deep-starvation target is close enough to balanced that the fixed prior is accidentally near-correct. Its value would not transfer to a target with a different marginal shape. Among genuine estimators the ordering MLLS, then JCPOT, BBSE and RLLS, then the kNN pre-pass is the meaningful one.

Dose-response, overlaid. Fig. 6 plots the coverage deciles of Fig. 8 with the best deployable estimator and with uniform balancing overlaid on the oracle. Both estimated curves keep the shape, the sign change and the cost at high coverage. The starved end loses about a quarter of the oracle amplitude, the break-even point moves from $r \approx 0.64$ to $r \approx 0.92$, and strict monotonicity is an oracle property: the estimated curves each contain one inversion, well inside overlapping error bars.

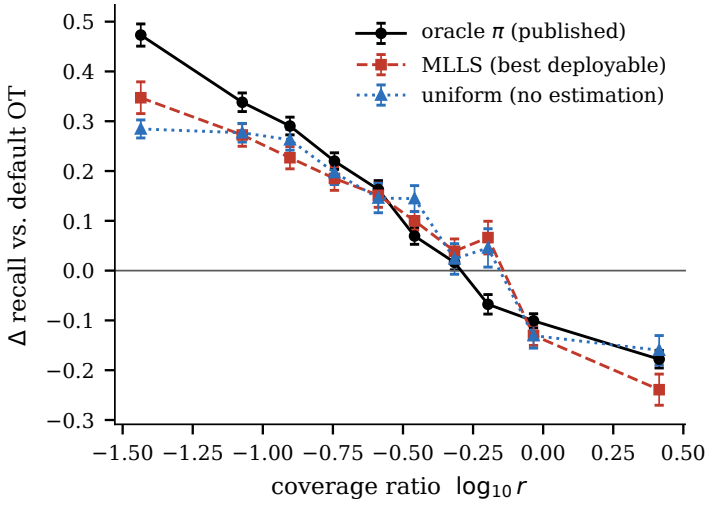


Figure 6: **Dose-response under a deployable class prior.** Coverage deciles as in Fig. 8, with the oracle marginal, the best deployable estimator and uniform balancing overlaid. The shape, the sign change and the cost at high coverage survive; the amplitude at the starved end and the break-even coverage do not.

Single-cell companion. The same analysis on the Janesick estimator panel ($n = 84$ class-observations, seven arms) is a companion rather than a replication, since the sample is nine times smaller and the two predictors remain correlated at $\rho(\log r, \text{separability}) = -0.167$. Two of the four relations replicate cleanly on all seven arms: removable is coverage-governed, at partial ρ from -0.41 to -0.73 with every arm at $p < 10^{-4}$, and residual is separability-governed, at partial ρ from -0.45 to -0.71 with every arm at $p < 2{\times}10^{-5}$. Two things break and we report them rather than smoothing them. First, removable depends on separability there even at the oracle ($\rho = +0.523$, $p = 3{\times}10^{-7}$), so the exact orthogonality of the second leg is specific to the constructed Office-Home design rather than a general property of the decomposition. Second, the residual's coverage leak is not detectable on any Janesick arm (p from 0.18 to 0.91), which at this sample size and with correlated predictors is an absence of evidence rather than evidence of absence. No single-cell table logs a separability column alongside all seven arms for the in-house cohort, so that version would require a re-run and is not reported.

G ADDITIONAL RESULTS

Table 13: Solver invariance and output-side detectors, single-cell (Jint). *Left:* four independent entropic-OT implementations on identical embeddings, points, labels, folds and seeds, with a kNN vote on the same embedding for reference. *Absorption* is the fraction of a starved class's target mass received by a single other class. Fixed-per-class target composition, so this is a within-run contrast across solvers rather than a cross-design comparison; the corrected natural-composition comparison is Table 15. *Right:* detectors for below-capacity points, 9 folds $\times$ 3 seeds, against a bar fixed in advance of selectivity at least $3\times$ at coverage at least 90%; suite verdict $0/5$, with $10/135$ unit evaluations passing. *over the units where it fires.

solver	macro	Endothelial	absorption	detector (side)	median sel. / pass
kNN vote (reference)	0.638	0.295	0.541	plan confidence (out)	2.43 / 6/27
UOT + centering	0.426	0.020	0.974	split conformal (out)	2.38 / 1/27
moscot	0.424	0.014	0.981	reject-sink OT (out)	1.51 / 0/27
UOT, no centering	0.320	0.009	0.994	transport entropy (out)	0.42 / 0/27
				coverage rule (source)	24.50^* / 3/27

Specificity on data with no composition shift. SpatialCorpus splits one file into source and target, so the two compositions agree by construction and a capacity correction should do nothing. It does nothing: uniform balancing leaves macro recall flat on all five tissues and lowers micro recall on every one. That is the coarse-grained counterpart of the sign change in Fig. 8, which shows the same specificity at class resolution.

Backbone invariance and cut-point sensitivity. Table 14 collects two robustness checks. On the left, repeating the mechanism analysis on a PCA basis reproduces every count, and PCA is the better read-out of the two, which forecloses the reading that PCA is simply worse; on PCA the source-side coverage rule fires in 24 of 27 units and passes 23, while every output-side statistic still fails. The same holds on Office-Home. Repeating the mechanism analysis on ResNet-50 followed by PCA gives a capacity partial of $+0.347$ against CLIP's $+0.329$, a distance partial of -0.191 against -0.206, and argmax counts of 20.2% nearest and 19.2% largest against 18.6% and 20.5% (13,312 pairs over 12 units). Both counts sit near the 1 in 7 a chance absorber would give, on both backbones, which is the same degeneracy the single-cell analysis shows and a second demonstration that the partial is the statistic that transfers. On the right, the below-capacity knob ordering of Table 3 holds at every cut point we tested, so the $r < 0.15$ boundary is a reporting convenience rather than a fitted one.

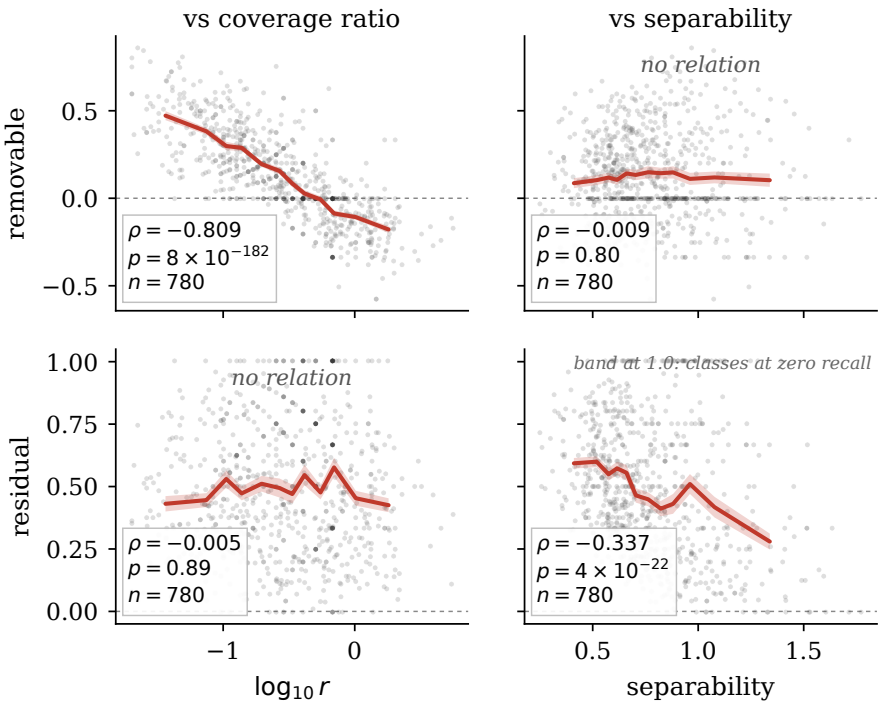
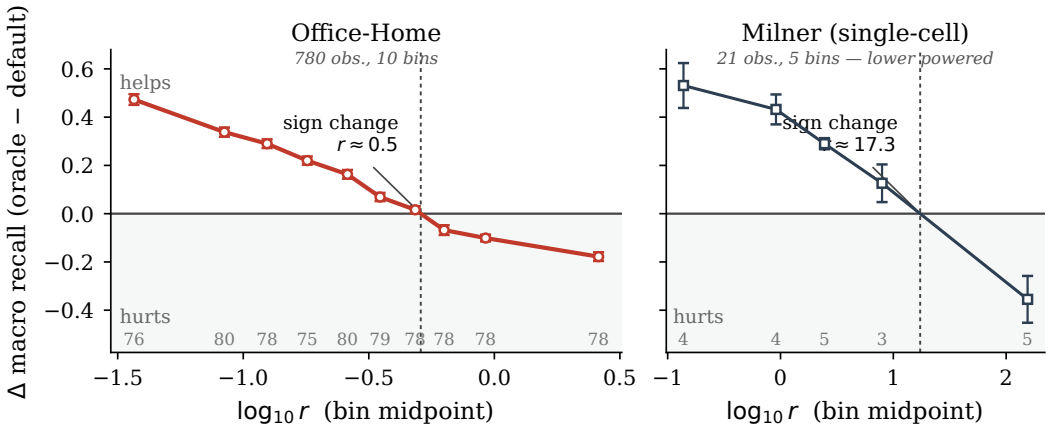


Figure 7: **The double dissociation.** What the composition-matched marginal removes (top) and what survives it (bottom), against coverage (left) and class separability (right), on Office-Home. The two null panels are shown deliberately: what is removed is explained by coverage and *not* by separability, and what survives is explained by separability and *not* by coverage. Points are class-observations (4 domain pairs $\times$ 3 seeds), the line is a binned mean with standard error, and the band at residual $= 1.0$ is classes the correction leaves at zero recall. The panels are interpretable only because the predictors were orthogonalized by construction; in the single-cell data they are confounded at $\rho = -0.618$.

Ratio, not sample size. A quantity stated as a ratio invites the objection that it is small-sample statistics in disguise. A trained classifier separates the two. In a fixed-per-class run, the most ratio-starved class ($r = 0.043$) has 117 effective training points and a trained latent-variable classifier (Xu et al., 2021) recovers it at 0.960 while transport loses it entirely, whereas a class four times better covered by ratio ($r = 0.439$) has only 54 training points and that classifier loses it, at 0.122. Transport follows the ratio and a trained classifier follows the count.

Pooled and stratified readings disagree. MLLS closes only 4.7% of the single-cell micro gap, the worst of the panel, while scoring 0.672 below capacity, above even the oracle's 0.546. That is possible because the oracle matches the *whole* composition rather than optimising the starved classes, so an estimator that over-allocates mass to rare classes can exceed it there and pay for it pooled, which is what MLLS does. Either number alone misleads, which is why every table in this paper carries both.



Grey row: class-observations per bin. Note the x-ranges differ — single-cell coverage extends two orders further, so the two crossing points are NOT comparable thresholds. The claim is the shape and the sign change, not a universal value of r.

Figure 8: **Dose-response of the correction.** Change in macro recall from the composition-matched marginal against coverage, binned. Monotone on both domains, crossing zero and turning negative on the best-covered classes. The single-cell panel has 21 class-observations in 5 bins against Office-Home's 780 in 10 and is a companion rather than an equally powered replication, and the x-ranges differ, so the two crossing points are not comparable thresholds.

Table 14: *Left:* backbone invariance, 9 folds $\times$ 3 seeds each on a 40,000-point source, fixed-per-class target, so a within-run contrast. Counts differ slightly from Sec. 3.2 because that run uses natural composition and this one does not. *Right:* the below-capacity knob ordering of Table 3 at four cut points, as Δ macro recall against the default marginal.

	foundation	PCA	cut	n	shape	τ	ε
confident errors	33/33	33/33	$r < 0.05$	56	0.472	0.243	0.006
absorber = nearest	0/33	0/33	$r < 0.10$	136	0.419	0.234	0.014
absorber = largest	30/33	33/33	$r < 0.15$	235	0.366	0.208	0.013
confidence ratio	0.874	0.906	$r < 0.30$	389	0.296	0.176	0.015
macro recall	0.445	0.495					
detector suite	0/5	0/5					

Estimator error is amplified, not merely passed through. A construction makes the failure mode concrete. Under a small *absolute* floor on estimated shares, a class the pre-pass never predicts holds share 0.0036 against 0.0766 for the rest, and its recall then stays exactly 0.000 even when its source points are a 20$\times$ better match to the target points in question, because the marginal starves the class, so it is never predicted, so the estimate is never contradicted. A *relative* floor of $0.5/K$ breaks the loop and is what every estimator arm reported here uses. The construction is illustrative rather than a measured arm, but the failure it describes is what we observed before the floor was changed, and the change is a one-line design rule for deploying Eq. 2: floor the estimate relative to uniform, not absolutely.

Table 15: Matched corrections, single-cell (Jint), 3 folds × 3 seeds, natural composition. *Composition-matched* arms, transport and vote alike, use the *true* target composition and are therefore matched oracles; *prior-corrected* arms divide by the source class prior and use no target information, so they are matched to each other at a lower information level. The oracle-to-oracle comparison is transport 0.546 against vote 0.414 below capacity, where *below* is the below-capacity stratum, three (fold, class) cells averaged over seeds.

read-out	macro	macro-F1	micro	below capacity
transport, default	0.433	0.322	0.393	0.030
kNN vote, uncorrected	0.627	0.553	0.625	0.351
linear probe, uncorrected	0.268	0.212	0.215	0.014
kNN, prior-corrected	**0.654**	0.571	0.640	0.430
kNN, composition-matched	0.635	0.577	0.652	0.414
linear probe, corrected	0.452	0.391	0.377	0.278
transport, composition-matched	0.619	**0.623**	**0.718**	**0.546**

Office-Home matched corrections, full. Table 16 is the Office-Home counterpart of Table 15. Two caveats travel with it. The linear-probe arms come from a separate Office-Home experiment with an independent composition draw, so differences of this size sit near what the comparison resolves; and the pooled metrics are dominated by well-covered classes, where Sec. 4.1 shows the marginal correction helps least and can hurt.

Table 16: Matched corrections on Office-Home, 4 domain pairs × 3 seeds, deep-starvation configuration. *Below* is the below-capacity stratum, $n = 235$ class-observations.

read-out	macro	macro-F1	micro	below capacity
transport, default	0.388	0.356	0.440	0.136
kNN vote, uncorrected	0.367	0.327	0.506	0.154
linear probe, uncorrected	0.199	0.136	0.313	0.004
kNN, prior-corrected	0.553	0.483	0.580	0.526
kNN, composition-matched	0.460	0.461	**0.644**	**0.566**
linear probe, class-balanced	**0.590**	**0.533**	0.641	0.542
linear probe, prior-corrected	0.389	0.329	0.517	0.182
transport, composition-matched	0.521	0.523	0.623	0.547

Which correction a simple read-out needs is domain-dependent. An uncorrected linear probe is catastrophic below capacity on both domains, at 0.014 and 0.004, and both corrections we tested lift it, but they rank oppositely. Dividing the posterior by the source prior reaches 0.278 on single-cell data where class-balanced loss re-weighting reaches 0.033, and on Office-Home the ordering reverses, at 0.542 against 0.182. We report both per domain and make no general recommendation. An uncorrected baseline understates a linear probe severely, which is a caution for any comparison of this kind, including those drawn here.

Janesick, natural composition. Table 17 gives macro and micro recall by marginal on both replicates. The pattern of Table 4 holds: uniform balancing takes the highest macro on rep2 while micro collapses, and only the composition-matched marginal raises macro without reducing micro on either replicate.

Per-pair replication, Office-Home. Oracle micro gap by domain pair: $+0.127$, $+0.174$, $+0.198$ and $+0.185$. It is present on all four, with the best estimator closing 53.7 to 70.0% on every one, so the estimator

Table 17: Janesick under natural target composition, macro / micro recall by marginal, two replicates $\times$ 3 seeds.

	replicate 1		replicate 2	
estimate of π	macro	micro	macro	micro
source frequencies	0.141	0.224	0.334	0.477
uniform	**0.206**	0.149	**0.591**	0.346
kNN pre-pass	0.180	0.167	0.479	0.437
JCPOT	0.179	0.157	0.461	0.389
BBSE	0.161	**0.316**	0.346	0.436
RLLS	0.160	0.317	0.347	0.436
MLLS	0.160	0.317	0.345	0.436
oracle π	0.204	0.314	0.449	**0.554**

contrast is not driven by a single pair. Which estimator is best is unstable across pairs, with MLLS at 70.0% on one and 15.0% on another, which is why we quote the range.

Minority synthesis does not add to the correction. Within-run contrasts on a fixed-per-class target. Synthesizing minority source points by latent-space interpolation (Chawla et al., 2002; Dablain et al., 2022), dosed to the fitted threshold, adds $+0.006$ macro recall on Jint and $+0.036$ on Janesick on its own, and $+0.000$ and $+0.002$ respectively on top of a corrected marginal, with absorption unchanged to three decimals. We also tested the one regime where it might have a niche, namely classes too small to reweight, with 17 and 24 target points, where it does not appear.

Impossibility cases. One target slide contains two classes with *zero* source points. These are absent rather than under-covered, since no marginal can place mass that does not exist, and they are excluded from all below-capacity means.