Forest-Guided Semantic Transport for Label-Supervised Manifold Alignment

Adrien Aumon[1]
adrien.aumon@umontreal.ca

Myriam Lizotte[1]
myriam.lizotte@mila.quebec

Guy Wolf[1]
wolfguy@mila.quebec

Kevin R. Moon[2]
kevin.moon@usu.edu

Jake S. Rhodes[3‡]
rhodes@stat.byu.edu

[1]Université de Montréal; Mila – Quebec AI Institute
[2]Utah State University [3]Brigham Young University
[‡]Corresponding author

Abstract

Label-supervised manifold alignment bridges the gap between unsupervised and correspondence-based paradigms by leveraging shared label information to align multimodal datasets. However, existing methods either rely on label-independent intra-domain geometry or incorporate supervision primarily through the alignment objective, rather than directly constructing task-aware intra-domain affinities. To address this limitation, we introduce FoSTA (Forest-guided Semantic Transport Alignment), which learns supervised forest geometries independently within each domain and uses their shared class-semantic structure to infer cross-domain correspondences without predefined anchors. FoSTA extends RF-GAP affinities to partially labeled data and aligns the resulting semantic representations through scalable hierarchical transport. Extensive comparisons with established baselines demonstrate strong correspondence recovery and semantic structure preservation on synthetic and real-world benchmarks, including multimodal data integration and single-cell batch correction.

1 Introduction

Manifold alignment (MA) seeks a shared representation for datasets collected in different modalities or feature spaces, and is a key primitive in multimodal learning and single-cell integration [1–3]. Most approaches fall into two broad families: *semi-supervised* methods that rely on a set of known cross-domain correspondences (anchors) [4, 5], and *unsupervised* methods that assume no shared information and instead match domains using geometric or distributional structure [6, 7]. In many practical settings, however, pointwise correspondences are impossible or expensive to obtain, while weaker side information—like shared class labels—is readily available (e.g., curated cell types across batches [8]). This motivates *label-supervised* manifold alignment, which uses labels to guide cross-domain matching without using anchors [9–11].

Most label-supervised alignment methods rely on Euclidean geometry to model intra-domain structure. This fails when only a subset of features is relevant to the task, as Euclidean distances weigh all coordinates equally, mixing signal with nuisance variation. In high-dimensional settings, this creates misleading neighborhoods that degrade the alignment. Tree ensemble affinities offer a principled alternative by adaptively emphasizing label-relevant structure. In particular, RF-GAP proximities [12] have been shown to outperform standard forest-based proximities and class-conditional Euclidean

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

kernels in guided representation learning [12–14]. However, existing forest-supervised alignment methods [15] use forest affinities for within-domain geometry while still requiring known cross-domain correspondences. Whether independently learned supervised forest geometries can instead provide sufficient shared structure to *infer* these correspondences remains unexplored.

To address this gap, we introduce **FoSTA** (Forest-guided Semantic Transport Alignment), a label-supervised framework that uses independently learned forest geometries to infer correspondences between heterogeneous domains without predefined anchors. We extend RF-GAP [12] to partial label assignment and use the resulting affinities to represent samples through their relationships to shared semantic classes. Cross-domain correspondences are then inferred by scalable semantic transport and propagated through the forest graphs to construct a joint geometry. The formulation further accommodates unequal domain sizes and partially overlapping class sets.

Across real-world multimodal benchmarks, FoSTA achieves strong correspondence recovery and label transfer compared to established baselines [9–11, 8]. On a single-cell batch alignment task, it achieves competitive batch correction while preserving biological variation. Controlled synthetic experiments further show its robustness to different domain transformations, levels of label availability, and design choices, confirming that the alignment gains mainly come from the upstream forest geometry.

2 Background

2.1 Guided manifold learning with tree ensembles

Manifold learning methods traditionally rely on unsupervised, Euclidean-based kernels (e.g., radial basis functions) to capture local geometric structure. When label information is available, however, meaningful representations should reflect not only feature–feature relationships but also feature–label dependencies. In such settings, purely unsupervised geometry can be misleading, as it treats all features equally regardless of their relevance to the label signal.

Pairwise proximities derived from leaf co-occurrence in trained Random Forests (RFs) or other tree ensembles [16–18, 12] address this limitation by jointly capturing similarity in the feature space and the task-relevant structure, implicitly downweighting features that are irrelevant for prediction. In particular, RF-GAP proximities [12] provide a robust label-informed proximity measure through a leaf weighting scheme based on in-bag/out-of-bag status, which reflects the predictive behavior of the underlying forest. This property has led to substantial improvements in guided representation learning compared to standard forest-based proximity definitions and Euclidean class-conditional approaches [13, 14]. In addition, tree-induced proximities naturally accommodate both continuous and categorical features and their exact calculation scales near-linearly in time and memory [19], whereas Euclidean kernels often require approximate nearest-neighbor methods to remain tractable [20].

2.2 Manifold alignment approaches

Given multiple domains containing observations from different modalities that describe related phenomena, manifold alignment aims to learn a shared representation that preserves both (i) the intrinsic geometric structure within each domain and (ii) information shared across domains. While intra-domain geometry can typically be inferred directly from the respective feature spaces, capturing shared cross-domain information depends critically on the assumptions made about how the domains are related.

Semi-supervised MA assumes access to a subset of one-to-one correspondences across domains, which serve as anchor points to guide alignment. These anchors provide explicit cross-domain constraints that allow the remaining unmatched samples to be embedded into a common space.

Traditional approaches include canonical correlation analysis [21], which learns linear projections that maximize the correlation between paired observations. While efficient, such global linear models are often inadequate for capturing the nonlinear structure of high-dimensional data. This limitation motivated subsequent graph-based methods [1, 2], which preserve local manifold geometry within each domain while enforcing alignment through anchor constraints. Alternatively, Wang and Mahadevan [4] projects both domains into a shared space using dimensionality reduction (DR), followed by Procrustes alignment on anchor points, and generally outperforms earlier approaches.

More recent semi-supervised MA methods leverage anchor correspondences while better preserving intra-domain geometry. Generative approaches [22] learn mappings between domains in a CycleGAN-style framework [23], augmented with correspondence losses to enforce alignment consistency and reduce superimposition. Duque et al. [5] incorporates diffusion geometry by using anchor correspondences to derive diffusion-informed inter-domain distances, which define the cost of an entropic OT problem [24]; the resulting coupling is then used to construct inter-domain affinities prior to joint embedding. In contrast, Rhodes and Rustad [25] aligns domains by directly linking their diffusion operators through a union graph. Building on this idea, Rhodes and Rustad [15] moves beyond Euclidean-based alignment by using RF-derived affinities within each domain. However, their method assumes fully labeled source and target domains, which could be restrictive; we relax this assumption in Section 3.2 by allowing partially labeled settings.

Other semi-supervised alignment methods also use twin autoencoders, incorporating geometry regularization in guided alignment [26] and supporting out-of-sample extension [27]. Related multimodal integration methods use other forms of cross-domain prior information. For instance, GLUE [28] uses a guidance graph encoding known relationships between features across modalities. While effective when such prior relationships are available, this assumption limits its applicability to general heterogeneous domains with unrelated feature spaces.

Anchor-based approaches rely on known correspondences, which can be expensive or impossible to obtain. This limitation is particularly pronounced in biological applications, where measurement processes are often destructive and prevent observing multiple modalities from the same cells.

Unsupervised MA is the most challenging setting, as it assumes no explicit shared information across domains. However, domains describing the same phenomenon are expected to exhibit similar underlying structure. This assumption allows correspondences to be inferred indirectly by relating points across domains based solely on their locations within their respective manifolds.

Early unsupervised approaches sought to recover shared structure by constructing joint representations from intra-domain geometry. Wang and Mahadevan [29] introduced a cross-domain heat kernel based on permutations of local geometric structure, combined with intra-domain kernels, followed by DR to obtain a shared embedding. Subsequent work framed unsupervised MA as a problem of implicit global distribution matching, e.g., by minimizing discrepancies between kernel mean embeddings [30] or by matching intrinsic geometric structure using geodesic distances rather than kernel matrices [31].

Rather than enforcing global matching, a complementary line of work infers correspondences through local neighborhood consistency. Haghverdi et al. [32] and Hie et al. [3] identify mutual nearest neighbors (NN) across domains and use these pairs as anchors to apply local, translation-based corrections. Similarly, Korsunsky et al. [33] and Welch et al. [34] rely on local alignment heuristics, using cluster-wise linear corrections or shared (dataset-invariant) latent factors, respectively.

To move beyond heuristic local matching and explicitly enforce global geometric consistency, another line of work formulates unsupervised MA as an OT problem. In this setting, correspondences are inferred globally using the Gromov–Wasserstein OT family (GWOT [6]), with discrepancies in geodesic distances serving as the transport cost, followed by DR [7, 8]. Alternatively, Chen et al. [35] extends Procrustes-based alignment by first projecting each domain into a shared latent space and then refining the alignment via Wasserstein Procrustes [36].

Even in the absence of explicit correspondences, one does not need to resort to fully unsupervised manifold alignment when side information is available in both domains.

Label-supervised MA occupies a middle ground between fully unsupervised and semi-supervised MA by leveraging an auxiliary form of shared information. A common source of such side information is a set of discrete class labels [9, 10, 37], for example, known cell types in single-cell analysis [8].

Early work in this setting directly learned a latent joint representation by incorporating label information into the alignment objective, enforcing strict same-class matching alongside intra-domain geometric regularization (KEMA [9, 10]). However, this strategy is highly sensitive to label availability, as samples with missing labels are left unmatched. In contrast, Manifold Alignment with Label Information (MALI [11]) adopts a two-stage approach in which cross-domain correspondences are first inferred through an OT formulation, and a joint similarity matrix is then constructed from the resulting coupling. This matrix can subsequently be used to obtain a shared latent representation if

desired. MALI proceeds similarly to Duque et al. [5], but replaces cross-domain diffusion-based distances through anchors with cosine distances between semantic representations, defined as projections onto a label space obtained by aggregating class-wise affinities within each domain.

In single-cell analysis, recent methods extend unsupervised GWOT alignment by incorporating labels into the transport cost, thus favoring matches between cells of the same type [8]. Others leverage label information through probabilistic deep generative models for batch correction [38]. However, these generative models typically assume that all domains share the same feature space, limiting them to homogeneous data integration.

These methods fall under a *weakly* supervised MA setting, as correspondences are inferred rather than provided explicitly, while still exploiting shared semantic information across domains. Existing label-supervised alignment methods differ in how supervision enters the alignment: MALI uses labels primarily to infer cross-domain correspondences, while KEMA incorporates class information directly into its alignment objective. In contrast, FoSTA uses labels to construct the intra-domain geometry itself through supervised forest affinities, which subsequently drive correspondence inference.

3 Methods

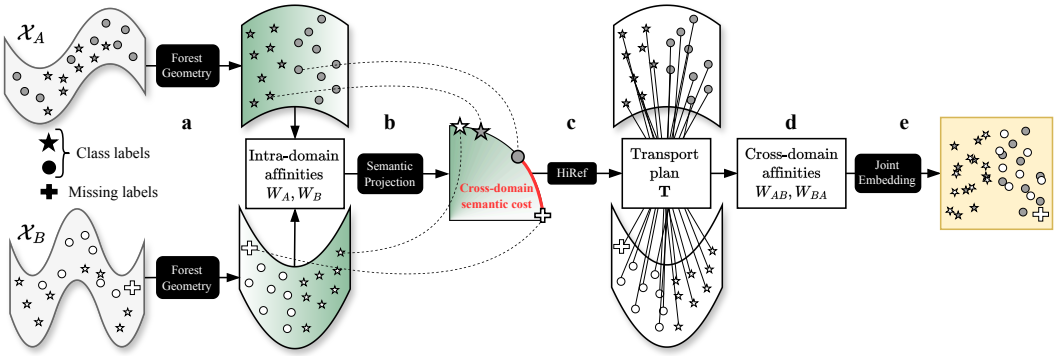

Figure 1: Overview of FoSTA. **(a)** Given two (partially labeled) domains $\mathcal{X}_A$ and $\mathcal{X}_B$, we compute semi-supervised, forest-derived affinities W_A and W_B to capture task-relevant neighborhood structure (Section 3.2). **(b)** We form class-wise semantic profiles and map samples from both domains to a shared, ℓ_2-normalized semantic space, which defines a cross-domain semantic cost (Section 3.3). **(c)** Using this cost, HiRef estimates an efficient bijective transport plan $\mathbf{T}$ between domains (Section 3.3). **(d)** We propagate correspondences through $\mathbf{T}$ to build cross-domain affinities W_{AB} and W_{BA} consistent with the intra-domain graphs (Section 3.4). **(e)** Finally, we compute a joint low-dimensional embedding from the resulting block affinity matrix (Section 3.5).

3.1 Problem formulation

Given a source domain $\mathcal{X}_A = \{\mathbf{x}_1^A, \ldots, \mathbf{x}_n^A\}$ and a target domain $\mathcal{X}_B = \{\mathbf{x}_1^B, \ldots, \mathbf{x}_m^B\}$, manifold alignment aims to learn a shared representation that preserves both intra-domain geometric structure and cross-domain relationships.

We consider a weakly supervised setting in which the two domains share a common label space, but explicit correspondences are unavailable. We further relax the fully labeled assumption by allowing both domains to be only partially labeled. Let $\mathcal{Y}_A = \{\mathbf{y}_1^A, \ldots, \mathbf{y}_r^A\}$ and $\mathcal{Y}_B = \{\mathbf{y}_1^B, \ldots, \mathbf{y}_s^B\}$, with $r \leq n$ and $s \leq m$, denote the observed labels in the source and target domains, respectively. We assume that domain labels take values in finite class sets $\mathcal{C}_A$ and $\mathcal{C}_B$, respectively, and that $\mathcal{C}_A \cap \mathcal{C}_B \neq \emptyset$. The goal is therefore to leverage this limited shared supervision to infer meaningful relationships across domains.

Our method focuses on learning cross-domain similarities that are comparable in scale to intra-domain affinities, which naturally supports both correspondence estimation and the construction of shared

embeddings [5, 11]. Thus, the core objective is to build a joint affinity matrix of the form

$$\mathbf{W} = \begin{bmatrix} W_A & W_{AB} \\ W_{BA} & W_B \end{bmatrix} \in \mathbb{R}_{\geq 0}^{(n+m)\times(n+m)},$$

where the diagonal blocks represent intra-domain affinities and the off-diagonal blocks represent cross-domain affinities. The construction of each affinity block is detailed in the sections that follow, and Fig. 1 presents a schematic representation of our FoSTA framework.

3.2 Label-informed intra-domain affinities

Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a partially labeled dataset, and denote by $\mathcal{I}^l$ and $\mathcal{I}^u$ the sets of labeled and unlabeled indices, respectively, with $\mathcal{I}^l \cup \mathcal{I}^u = \{1, \ldots, n\}$. Our goal is to construct class-aware intra-domain affinities that better reflect the label structure than standard distance-based kernels.

We train a bootstrapped forest classifier (typically an RF) on the labeled subset $\{\mathbf{x}_i : i \in \mathcal{I}^l\}$ and use it to define affinities between all pairs of points. Let $\mathcal{T}$ be the set of trees, with $T = |\mathcal{T}|$. For a query point $\mathbf{x}_i$, denote by $S_i \subset \mathcal{T}$ the set of trees in which $\mathbf{x}_i$ is out-of-bag. For each tree $t \in \mathcal{T}$, let $\ell_i(t)$ be the leaf reached by $\mathbf{x}_i$, let $M_i(t)$ be the multiset of in-bag training samples that fall in that leaf, and let $c_j(t)$ denote the bootstrap multiplicity of any labeled point $\mathbf{x}_j$. We define the intra-domain affinity matrix W by

$$W_{ij} = \frac{1}{|S_i|} \sum_{t \in S_i} \frac{c_j(t)\, \mathbb{I}(\ell_i(t) = \ell_j(t))}{|M_i(t)|} \in \mathbb{R}_{\geq 0}^{n\times n}. \tag{1}$$

When the target index $j \in \mathcal{I}^l$, this expression coincides with the standard RF-GAP proximity [12]. For unlabeled points, however, the bootstrap multiplicity $c_j(t)$ is not available, since these samples are not used during training. To extend the definition, we treat each unlabeled point as a *phantom training sample*: it is routed through the forest exactly as a labeled sample, but its contribution on the target side is replaced by a simple deterministic surrogate derived from labeled bootstrap statistics [39]. Concretely, unlabeled points act as out-of-bag queries while contributing as unit-weight in-bag targets, i.e., $c_j(t) = 1$ for $j \in \mathcal{I}^u$. This choice preserves the scale of RF-GAP contributions while extending the weighted leaf co-occurrence mechanism to unlabeled targets. We further justify this surrogate by showing that it provides an optimal approximation, in a least-squares sense, of the labeled multiplicities under standard RF assumptions (Lemma 1).

3.3 Learning correspondences via semantic transport

For clarity, we focus on the balanced setting with $n = m$ and assume full semantic overlap between the source and target domains, i.e. $\mathcal{C}_A = \mathcal{C}_B = \{1, \ldots, C\}$. Extensions to imbalanced settings and partially overlapping domains are discussed in Appendix B. To establish correspondences across domains, we seek a shared representation that abstracts away domain-specific variability while preserving class-level structure. Starting from the label-informed RF-GAP affinity matrix W_* in domain $* \in \{A, B\}$ (Eq. 1), we construct a *semantic profile* for each sample that summarizes its relationships to labeled data across classes. Specifically, for any point $\mathbf{x}_i^* \in \mathcal{X}_*$, we define a C-dimensional semantic vector

$$\tilde{S}_*(i,:) \;=\; \frac{S_*(i,:)}{\|S_*(i,:)\|_2}, \qquad S_*(i,c) \;=\; \frac{1}{p_c^*} \sum_{j \in \mathcal{I}_*^c} (W_*)_{ij}. \tag{2}$$

Here, $\mathcal{I}_*^c$ denotes the set of samples in domain $*$ labeled with class c, and $p_c^* = |\mathcal{I}_*^c|/|\mathcal{I}_*^l|$ is the estimated class prior. The factor $1/p_c^*$ mitigates class imbalance, while the ℓ_2 normalization makes the representation scale-invariant, so alignment depends on the relative shape of the class-affinity profile rather than its total magnitude. Each sample is therefore embedded into a shared semantic space, regardless of its original domain. Because we aggregate affinities only to labeled samples (Eq. 2), the unlabeled–unlabeled portion of W_* does not define the label-supervised semantic signal. Instead, it helps preserve a coherent affinity structure for the downstream joint embedding once labeled and unlabeled samples are embedded together. Intuitively, samples with similar semantic roles are mapped close to each other, while samples associated with different classes are separated. Unlike MALI, which applies DPT to an unsupervised geometry before semantic matching, FoSTA derives semantic representations directly from its supervised forest geometry.

To align samples across domains, one could naively rely on a cross-domain k-NN graph. However, formulating alignment as an OT assignment has been shown to provide greater robustness by enforcing global structural consistency across domains [5, 11]. Given normalized semantic representations $\tilde{S}_A, \tilde{S}_B \in \mathbb{R}^{n \times C}$, we define the OT problem

$$\mathbf{T}^\star = \arg\min_{\mathbf{T} \in \mathcal{U}} \langle \mathbf{T}, \mathbf{C} \rangle, \qquad \mathcal{U} = \left\{ \mathbf{T} \in \{0,1\}^{n \times n} \;\middle|\; \mathbf{T}\mathbf{1} = \mathbf{1},\ \mathbf{T}^\top \mathbf{1} = \mathbf{1} \right\},$$

where $\mathbf{C}_{ij} = \|\tilde{S}_A(i,:) - \tilde{S}_B(j,:)\|_2^2$ is the pairwise cost matrix, and $\langle \cdot, \cdot \rangle$ denotes the Frobenius inner product. The optimal coupling $\mathbf{T}^\star$ defines a one-to-one correspondence between samples in the two domains. Solving this OT problem exactly is computationally prohibitive for large-scale datasets, as it requires explicit construction of the dense cost matrix $\mathbf{C}$, leading to at least quadratic time and memory complexity. We therefore approximate $\mathbf{T}^\star$ using HiRef [40], which recursively partitions the two domains and solves a sequence of smaller, localized transport problems. This divide-and-conquer strategy yields quasilinear complexity while producing a sparse, bijective Monge map.

Finally, note that on the unit hypersphere induced by ℓ_2 normalization in Eq. 2, minimizing squared Euclidean distance is equivalent to maximizing cosine similarity (Lemma 2). This equivalence allows HiRef to leverage fast low-rank factorizations of the Euclidean cost [41] without explicitly forming $\mathbf{C}$, effectively performing cosine-based semantic matching while retaining the scalability of the Euclidean-based HiRef solver. The resulting coupling $\mathbf{T} \approx \mathbf{T}^\star$ thus provides an efficient and robust cross-domain correspondence.

3.4 Cross-domain affinities through correspondence propagation

While one could directly populate the off-diagonal blocks of the joint affinity matrix using the binary assignment encoded by $\mathbf{T}$ (i.e., setting entries to either 0 or 1), such a construction fails to capture meaningful similarities between pairs that are not directly matched. That is, it does not provide a notion of graded cross-domain similarity beyond the hard assignment.

To obtain smoother and more informative cross-domain affinities, we propagate intra-domain affinities in Eq. 1 across the estimated correspondences by transferring local geometric structure from each domain to the other:

$$W_{AB} = W_A \mathbf{T}, \qquad W_{BA} = W_B \mathbf{T}^\top.$$

As a result, two samples from different domains are assigned a high affinity if they are either directly matched by $\mathbf{T}$ or if they are close to matched samples within their respective domains. This construction yields cross-domain similarities that are consistent in scale with intra-domain affinities and better reflect the underlying manifold structure across domains.

3.5 Joint embedding

The final step of our framework projects both domains into a shared latent space. Given the label-informed cross-domain affinities, a natural approach would be to perform a barycentric projection, i.e., to map samples from one domain onto the other using weighted averages of neighboring coordinates. However, such projections are highly sensitive to the curse of dimensionality and operate directly in the original feature space. This is undesirable in our setting, as the raw feature space may be noisy and may not faithfully reflect the underlying phenomenon of interest. Instead, we rely on nonlinear, diffusion-based DR to embed both domains into a shared low-dimensional representation. Specifically, we use Landmark PHATE [42], a scalable variant of PHATE, which has been shown to work effectively with forest-derived affinities in guided representation learning [13, 14].

3.6 Complexity analysis

FoSTA is designed to scale efficiently with the number of samples. Its sparse leaf-based RF-GAP affinities, hierarchical OT, and Landmark PHATE avoid dense pairwise computations, resulting in near-linear time and memory complexity. A detailed complexity analysis is provided in Appendix D, with empirical scaling results reported in Appendix E.

4 Empirical results

Code to reproduce all experiments is publicly available at `https://github.com/aumona/FoSTA`.

Table 1: **Model performance on real-world multimodal alignment benchmarks.** Results are averaged over five seeds for label transfer accuracy (Acc), alignment score (AS), and FOSCTTM (FOS), with ↑/↓ indicating better performance. Gold, silver, and bronze denote the top three methods. RGB-D and Sketchy include both ResNet18 [48] and DINOv2-Base [49] features. Top-k accuracy (Acc@5 or Acc@10) is also reported where applicable. RGB-D scores are not reported for Pamona due to excessive runtimes (> 10 hours).

Alignment task	Metric	**FoSTA**	KEMAlin	KEMArbf	MALI	Pamona
Audio ↔ Video (AVE, 28 classes)	Acc↑	**0.404**±0.01	0.226±0.01	**0.362**±0.02	**0.374**±0.01	0.202±0.01
	Acc@5↑	**0.786**±0.01	0.687±0.01	**0.747**±0.00	**0.765**±0.01	0.442±0.02
	AS↑	0.914±0.00	**0.921**±0.00	**0.930**±0.00	**1.002**±0.00	0.333±0.01
	FOS↓	**0.093**±0.00	**0.167**±0.00	**0.075**±0.00	0.170±0.00	0.196±0.02
Image ↔ Depth crop (RGB-D DINOv2-B, 51 classes)	Acc↑	**0.324**±0.00	**0.287**±0.00	**0.191**±0.02	0.097±0.02	—
	Acc@10↑	**0.792**±0.00	**0.813**±0.00	**0.578**±0.05	0.469±0.03	—
	AS↑	**0.934**±0.01	**0.944**±0.00	0.884±0.03	**0.984**±0.00	—
	FOS↓	**0.079**±0.00	**0.082**±0.00	**0.133**±0.02	0.294±0.02	—
Image ↔ Depth crop (RGB-D ResNet18, 51 classes)	Acc↑	**0.284**±0.01	**0.213**±0.00	**0.208**±0.03	0.085±0.01	—
	Acc@10↑	**0.813**±0.00	**0.768**±0.01	**0.585**±0.04	0.431±0.02	—
	AS↑	**0.936**±0.00	**0.972**±0.00	0.898±0.02	**0.986**±0.00	—
	FOS↓	**0.080**±0.00	**0.112**±0.00	**0.124**±0.01	0.342±0.01	—
Image ↔ Human sketch (Sketchy DINOv2-B, 125 classes)	Acc↑	**0.137**±0.01	0.092±0.01	**0.209**±0.02	**0.106**±0.00	0.011±0.00
	Acc@10↑	**0.613**±0.01	**0.521**±0.03	**0.596**±0.01	0.410±0.01	0.117±0.01
	AS↑	**0.950**±0.00	**0.947**±0.01	0.934±0.01	**0.980**±0.00	0.392±0.02
	FOS↓	**0.114**±0.00	**0.130**±0.01	**0.061**±0.00	0.246±0.01	0.392±0.02
Image ↔ Human sketch (Sketchy ResNet18, 125 classes)	Acc↑	**0.096**±0.00	0.045±0.00	**0.090**±0.00	**0.151**±0.01	0.021±0.00
	Acc@10↑	**0.452**±0.00	0.315±0.01	**0.385**±0.01	**0.504**±0.01	0.204±0.01
	AS↑	**0.959**±0.00	**0.866**±0.00	0.840±0.02	**0.987**±0.00	0.735±0.03
	FOS↓	**0.146**±0.00	0.209±0.00	**0.134**±0.00	**0.173**±0.00	0.342±0.02
Sensor 1 ↔ Sensor 2 (HAR, 6 classes)	Acc↑	**0.627**±0.01	0.215±0.00	**0.506**±0.01	**0.619**±0.01	0.139±0.02
	AS↑	**0.681**±0.02	0.081±0.00	**0.847**±0.03	**0.891**±0.01	0.108±0.01
	FOS↓	**0.175**±0.00	0.479±0.00	**0.195**±0.01	**0.243**±0.00	0.435±0.05

4.1 Multimodal alignment for data integration

We compare FoSTA with four label-supervised manifold alignment baselines: kernel manifold alignment with radial basis function (KEMArbf [10]) and its linear variant (KEMAlin [9]), MALI [11], and Pamona [8]. As our evaluation focuses on joint visualization, we use two-dimensional embeddings for all methods. Other methods discussed in Section 2.2 are not included because they are either fully unsupervised or assume additional cross-domain information, such as known anchors or shared feature spaces. Implementation and hyperparameter settings are provided in Appendix F.

We consider four real-world multimodal alignment tasks in which paired domains provide distinct views of the same underlying observations: audio and video from AVE [43], RGB and depth images from RGB-D [44], photos and human sketches from Sketchy [45], and accelerometer and gyroscope measurements from HAR [46]. We evaluate the methods under partial label supervision by masking a subset of labels during alignment and evaluating label transfer on the held-out labels. Table 1 reports three complementary criteria: label-transfer accuracy [31, 7], domain mixing through the Alignment Score (AS [47]), and correspondence recovery through the Fraction of Samples Closer Than the True Match (FOSCTTM [30]). Full experimental details are provided in Appendix G.

Across the four real-world tasks, FoSTA shows consistently strong label transfer and correspondence recovery across modalities and feature representations. It achieves the best accuracy on AVE, both RGB-D feature representations, and HAR, as well as the best top-k accuracy on AVE, RGB-D with ResNet18, and Sketchy with DINOv2-B. FoSTA also achieves the best FOSCTTM on both RGB-D

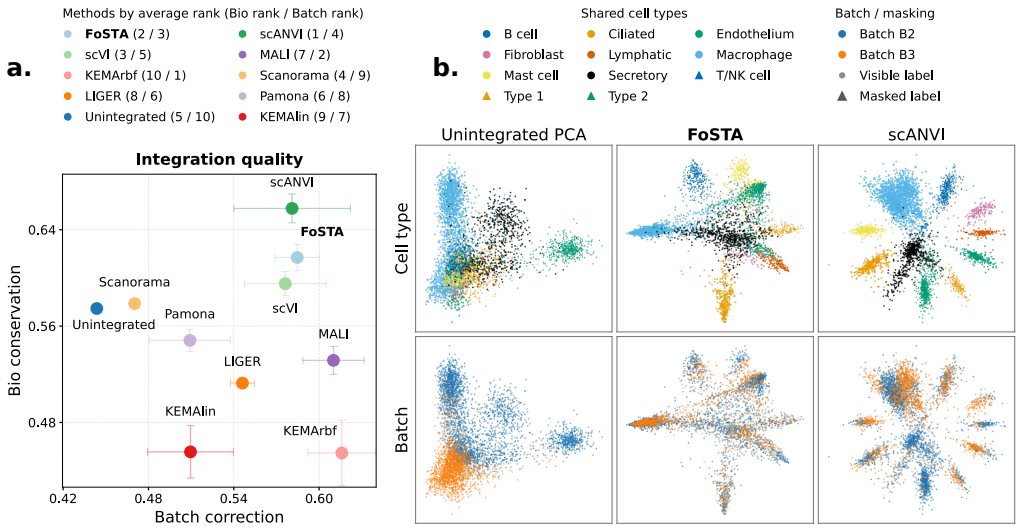

Figure 2: **Results of batch effect removal on lung single-cell data with 20% label masking [50].** **a.** Biological preservation (Bio) vs. batch correction (Batch) scores [51], averaged across 36 batch pairs and five seeds. Error bars represent seed variability, computed as the average within-pair standard deviation across the 36 batch pairs. **b.** Embeddings of the B2/B3 batch integration task. Left column: data before correcting for batch effects, visualized with PCA. Center column: resulting embedding after applying FoSTA. Right column: resulting embedding after applying scANVI. The plots are colored by cell type (top) and batch (bottom). FoSTA mixes the batches well while preserving the biological structure of the data.

representations and HAR, while remaining competitive on Sketchy. Other methods perform better on individual criteria, notably KEMArbf for correspondence recovery on AVE and Sketchy and MALI for AS on several datasets. FoSTA nevertheless remains robust across alignment settings, consistently ranking among the top two methods in both label-transfer accuracy and FOSCTTM, and among the top three in AS in all settings except AVE. As further discussed in Appendix H, we place greater emphasis on label transfer and FOSCTTM, which directly measure semantic preservation and ground-truth correspondence recovery, respectively, rather than domain mixing alone. The UCI experiments (Appendices H.1 and H.2) complement these real-world results by isolating different forms of cross-domain perturbation, where FoSTA achieves consistently strong accuracy and FOSCTTM across most settings. Finally, the ablations in Appendices H.3 and H.4 show that these gains are robust to secondary design choices and partial label availability, and are primarily driven by the upstream forest geometry.

4.2 Batch effect removal in single-cell genomics

We now consider the task of batch effect removal in single-cell genomics, where the goal is to combine data batches of a single modality (i.e., sharing the same feature space). Different batches may contain data collection artifacts called batch effects, which are due to differences in environment or experimental setup (e.g., protocol, time, instrument, or donors). Batch effects are widespread in single-cell data and must be removed so that downstream analyses are not impacted by them [32, 52, 53, 51, 54].

We use the lung atlas dataset from Vieira Braga et al. [50], following the benchmark setting of Luecken et al. [51]. We evaluate 36 real within-family batch pairs, which may contain only partially overlapping cell types. To evaluate the methods under partial label supervision, we independently mask 20% of the cell-type labels in each batch during alignment. Full dataset, preprocessing, and masking details are provided in Appendix I.1.

We compare FoSTA with eight prominent batch correction and manifold alignment methods. These include Scanorama [3] and LIGER [34], which are unsupervised methods relying on local anchoring or matrix-factorization strategies; scVI and its label-supervised variant scANVI [38], which are prob-

abilistic deep generative models; MALI [11] and Pamona [8], which are label-supervised manifold alignment methods based on optimal transport; and the linear and RBF variants of KEMA [9, 10], which are label-supervised kernel-based approaches. Following Luecken et al. [51], we evaluate biological conservation and batch correction. We note that these metrics differ from those used in Section 4.1, since there are no known cell-level correspondences between batches. Full evaluation details are provided in Appendix I.2, implementation settings in Appendix F, and complete quantitative results in Appendix I.3.

As discussed in Luecken et al. [51], batch correction can come at the expense of biological conservation by removing meaningful biological variation associated with batch differences. Our results follow this pattern (Fig. 2a). With the exception of FoSTA, scVI, scANVI, and, to a lesser extent, Scanorama, methods that improve batch correction over the unintegrated representation generally achieve lower biological conservation. FoSTA achieves a strong balance between the two objectives, ranking second in biological conservation and third in batch correction, and tying with scANVI for the best overall average rank. scANVI achieves the highest biological conservation, while FoSTA provides stronger batch correction. Conversely, MALI and KEMArbf achieve high batch correction but substantially lower biological conservation. This balance remains robust under reduced label availability, with FoSTA maintaining strong biological conservation and batch correction when 50% of the labels are masked (Appendix I.4). Overall, FoSTA remains competitive with methods specifically designed for single-cell batch correction such as scVI/scANVI while also supporting the heterogeneous feature spaces considered in our multimodal alignment experiments.

To complement the aggregate metrics, Fig. 2b shows a representative B2/B3 batch pair. Beyond reducing batch separation, FoSTA reveals continuous structure between cell populations that is consistent with known lung organization. In particular, the endothelial populations (Endothelium/BEC and Lymphatic/LEC) and the stromal population (Fibroblast) form a connected neighborhood, with LECs at its edge. This captures both the relationship between blood and lymphatic endothelium and their association with fibroblasts [55]. In scANVI, these populations form more isolated clusters, with LECs separating BECs from fibroblasts.

FoSTA also places the epithelial Secretory population at the center of a broader structure, connecting the BEC–LEC–Fibroblast neighborhood to other epithelial populations (Ciliated, Type 1, and Type 2). This agrees with known interactions between epithelial, endothelial, and stromal populations in the lung [56], as well as established relationships between secretory, ciliated, and alveolar cells [57, 58]. Under scANVI, Secretory instead forms a more isolated branch, making these relationships less apparent. Thus, FoSTA does more than separate cell types and mix batches: it preserves continuous biological organization between cell populations that is less apparent in the more specialized scANVI representation. Additional visualizations including scVI, MALI, and KEMArbf are provided in Appendix I.3.

Overall, FoSTA achieves a strong balance between biological conservation and batch correction while recovering meaningful class-aware biological structure. Although methods such as scVI and scANVI are specifically designed for single-cell integration, FoSTA remains competitive while also applying to heterogeneous modalities with unrelated feature spaces.

5 Conclusion and discussion

We proposed FoSTA, a forest-guided framework for label-supervised manifold alignment that learns task-aware geometry independently within each domain and uses this shared semantic structure to infer cross-domain correspondences without predefined anchors. By extending RF-GAP affinities to partially labeled data and combining them with scalable semantic transport, FoSTA provides a general framework for aligning domains with potentially unrelated feature spaces.

Across real-world multimodal benchmarks spanning audio–video, RGB–depth, photo–sketch, and sensor modalities, FoSTA achieves strong label transfer and correspondence recovery compared with established weakly supervised alignment methods. On real single-cell batch alignment tasks, FoSTA also achieves a strong balance between biological conservation and batch correction, remaining competitive with specialized methods such as scVI and scANVI. The resulting embeddings preserve continuous relationships between biologically related cell populations that are less apparent in more strongly clustered representations. Together with our controlled perturbation experiments and

ablations, these results show that FoSTA's alignment gains are primarily driven by the upstream forest geometry.

Beyond its empirical performance, FoSTA accommodates partial label availability, unequal domain sizes, and partially overlapping class sets, while its sparse forest affinities and hierarchical transport provide favorable scaling with the number of samples. Its main limitations remain its dependence on the quality of the underlying forest geometry and the current restriction to pairwise alignment. Future work could improve robustness to challenging transformations through modern oblique forests and extend the framework to multiple domains. Overall, our results establish task-aware forest geometry as an effective foundation for weakly supervised alignment across heterogeneous multimodal data and specialized applications such as single-cell batch correction.

AI use statement

In this work, we used generative AI tools to assist with method implementation and the development of experimental and result-compilation pipelines. We additionally used generative AI tools to assist with drafting and polishing parts of the manuscript. All AI-assisted code was reviewed and tested by the authors, and all reported results were independently verified. We reviewed all AI-assisted work and take responsibility for the final content of this work, including text, code, and artifacts produced with the aid of generative AI.

Broader impacts

FoSTA provides a general framework for aligning heterogeneous datasets using shared semantic annotations without requiring known sample correspondences or shared feature spaces. This may facilitate the joint analysis and visualization of complementary data sources that are otherwise difficult to compare, with potential applications in multimodal and biomedical data analysis. In particular, its ability to operate with partial label information and partially overlapping classes may be useful in settings where annotations are costly or incomplete, such as single-cell genomics and other biological applications.

Like many machine learning pipelines operating on sensitive biological or clinical measurements, FoSTA may inherit privacy and fairness concerns from the underlying data and annotations. In particular, biomedical datasets may contain patient-derived information whose use requires appropriate consent, anonymization, and secure handling. Moreover, because FoSTA uses class labels to guide the learned geometry, biases or inaccuracies in these annotations may be reflected in the resulting alignment and representation. These concerns are not unique to FoSTA but should be considered when applying supervised representation learning to sensitive data. Future work could explore privacy-preserving extensions and robustness to demographic, annotation, or sampling biases in partially labeled biomedical datasets.

Acknowledgments and Disclosure of Funding

This research was supported in part by the Ministry of Health and Social Services (Quebec) in collaboration with the Centre intégré de santé et de services sociaux Centre-Sud-de-l'Île-de-Montréal [A.A.], the Outstanding PhD Candidate Scholarship from the Institut des sciences mathématiques [A.A.], FRQNT Doctoral research scholarship [M.L.], Canada CIFAR AI Chair [G.W.], NSERC Discovery grant 03267 [G.W.], FRQNT-NSERC grant 2023-NOVA-329125 [G.W.], NSF grant DMS-2327211 [G.W.], NSF grant 221232 [K.M], NIH grant 1R15HL168697 [K.M.], and the IVADO Visiting Scholar program [K.M]. The authors are solely responsible for the content of this work, and the views expressed in it do not necessarily reflect those of the funding agencies. This research was enabled in part by compute resources provided by Mila and Université de Montréal (incl. Compute Canada and Calcul Quebec). We sincerely thank the anonymous reviewers and Area Chairs for their constructive feedback and efforts throughout the review process, which helped us improve the paper.

References

[1] Ji Hun Ham, Daniel D. Lee, and Lawrence K. Saul. Learning high dimensional correspondences from low dimensional manifolds. In *Proceedings of the ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*, pages 34–39, 2003.

[2] Jihun Ham, Daniel Lee, and Lawrence Saul. Semisupervised alignment of manifolds. In *International workshop on artificial intelligence and statistics*, pages 120–127. PMLR, 2005.

[3] Brian Hie, Bryan Bryson, and Bonnie Berger. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nature Biotechnology*, 37(6):685–691, June 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0113-3. tex.copyright: 2019 The Author(s), under exclusive licence to Springer Nature America, Inc.

[4] Chang Wang and Sridhar Mahadevan. Manifold alignment using Procrustes analysis. In *Proceedings of the 25th international conference on Machine learning - ICML '08*, pages 1120–1127, Helsinki, Finland, 2008. ACM Press. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390297.

[5] Andrés F. Duque, Guy Wolf, and Kevin R. Moon. Diffusion Transport Alignment. In Bruno Crémilleux, Sibylle Hess, and Siegfried Nijssen, editors, *Advances in Intelligent Data Analysis XXI*, pages 116–129, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-30047-9. doi: 10.1007/978-3-031-30047-9_10.

[6] Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Foundations of computational mathematics*, 11(4):417–487, 2011.

[7] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. SCOT: Single-cell multi-omics alignment with optimal transport. *Journal of computational biology*, 29(1): 3–18, 2022.

[8] Kai Cao, Yiguang Hong, and Lin Wan. Manifold alignment for heterogeneous single-cell multi-omics data integration using Pamona. *Bioinformatics (Oxford, England)*, 38(1):211–219, December 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab594.

[9] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI proceedings-international joint conference on artificial intelligence*, volume 22, page 1541, 2011.

[10] Devis Tuia and Gustau Camps-Valls. Kernel Manifold Alignment for Domain Adaptation. *PLOS ONE*, 11 (2):e0148655, February 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0148655.

[11] Andres F. Duque, Myriam Lizotte, Guy Wolf, and Kevin R. Moon. Manifold Alignment with Label Information. In *2023 International Conference on Sampling Theory and Applications (SampTA)*, pages 1–6, July 2023. doi: 10.1109/SampTA59647.2023.10301369.

[12] Jake S. Rhodes, Adele Cutler, and Kevin R. Moon. Geometry- and Accuracy-Preserving Random Forest Proximities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10947–10959, September 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2023.3263774.

[13] Jake S. Rhodes, Adrien Aumon, Sacha Morin, Marc Girard, Catherine Larochelle, Elsa Brunet-Ratnasingham, Amélie Pagliuzza, Lorie Marchitto, Wei Zhang, Adele Cutler, François Grand'Maison, Anhong Zhou, Andrés Finzi, Nicolas Chomont, Daniel E. Kaufmann, Stephanie Zandee, Alexandre Prat, Guy Wolf, and Kevin R. Moon. Gaining biological insights through supervised data visualization. *Nature Computational Science*, 6(7):732–752, July 2026. ISSN 2662-8457. doi: 10.1038/s43588-026-00999-7. URL `https://doi.org/10.1038/s43588-026-00999-7`.

[14] Adrien Aumon, Shuang Ni, Myriam Lizotte, Guy Wolf, Kevin R. Moon, and Jake Slater Rhodes. Random Forest Autoencoders for Guided Representation Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025.

[15] Jake S. Rhodes and Adam G. Rustad. Random Forest-Supervised Manifold Alignment. In *2024 IEEE International Conference on Big Data (BigData)*, pages 3309–3312, December 2024. doi: 10.1109/BigData62323.2024.10825663.

[16] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

[17] Leo Breiman and Adele Cutler. Manual on setting up, using, and understanding random forests v4.0. Technical report, University of California, Berkeley, Department of Statistics, 2003.

[18] Sarah Tan, Matvey Soloviev, Giles Hooker, and Martin T. Wells. Tree Space Prototypes: Another Look at Making Tree Ensembles Interpretable. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, FODS '20, pages 23–34, New York, NY, USA, October 2020. Association for Computing Machinery. ISBN 978-1-4503-8103-1. doi: 10.1145/3412815.3416893.

[19] Adrien Aumon, Guy Wolf, Kevin R. Moon, and Jake S. Rhodes. Revisiting forest proximities via sparse leaf-incidence kernels, 2026. URL `https://arxiv.org/abs/2601.02735`. arXiv: 2601.02735 [cs.LG].

[20] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, 1998.

[21] Bruce Thompson. *Canonical correlation analysis: Uses and interpretation*, volume 47. Sage, 1984.

[22] Matthew Amodio and Smita Krishnaswamy. MAGAN: Aligning Biological Manifolds. In *Proceedings of the 35th International Conference on Machine Learning*, pages 215–223. PMLR, July 2018.

[23] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[24] Gabriel Peyré, Marco Cuturi, and others. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[25] Jake S. Rhodes and Adam G. Rustad. Graph Integration for Diffusion-Based Manifold Alignment. In *2024 International Conference on Machine Learning and Applications (ICMLA)*, pages 1–8, December 2024. doi: 10.1109/ICMLA61862.2024.00026.

[26] Jake S. Rhodes, Adam G. Rustad, Marshall S. Nielsen, Morgan Chase McClellan, Dallan Gardner, and Dawson Hedges. Guided manifold alignment with geometry-regularized twin autoencoders. In *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1454–1463, 2025. doi: 10.1109/ICDMW69685.2025.00174.

[27] Jake S. Rhodes, Adam G. Rustad, and Marshall S. Nielsen. Enabling out-of-sample extension in semi-supervised manifold alignment through twin autoencoders. In *2025 International Conference on Sampling Theory and Applications (SampTA)*, pages 1–5, 2025. doi: 10.1109/SampTA64769.2025.11133509.

[28] Zhi-Jie Cao and Ge Gao. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nature Biotechnology*, 40(10):1458–1466, October 2022. ISSN 1546-1696. doi: 10.1038/s41587-022-01284-4. URL `https://www.nature.com/articles/s41587-022-01284-4`.

[29] Chang Wang and Sridhar Mahadevan. Manifold alignment without correspondence. In *IJCAI*, volume 2, page 3, 2009.

[30] Jie Liu, Yuanhao Huang, Ritambhara Singh, Jean-Philippe Vert, and William Stafford Noble. Jointly Embedding Multiple Single-Cell Omics Measurements. *Algorithms in bioinformatics : ... International Workshop, WABI ..., proceedings. WABI (Workshop)*, 143:10, September 2019. doi: 10.4230/LIPIcs.WABI.2019.10.

[31] Kai Cao, Xiangqi Bai, Yiguang Hong, and Lin Wan. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics (Oxford, England)*, 36(Supplement_1):i48–i56, July 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa443.

[32] Laleh Haghverdi, Aaron T. L. Lun, Michael D. Morgan, and John C. Marioni. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36 (5):421–427, May 2018. ISSN 1546-1696. doi: 10.1038/nbt.4091. tex.copyright: 2018 Springer Nature America, Inc.

[33] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296, December 2019. ISSN 1548-7105. doi: 10.1038/s41592-019-0619-0. tex.copyright: 2019 The Author(s), under exclusive licence to Springer Nature America, Inc.

[34] Joshua D. Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, and Evan Z. Macosko. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell*, 177(7):1873, June 2019. doi: 10.1016/j.cell.2019.05.006.

[35] Dexiong Chen, Bowen Fan, Carlos Oliver, and Karsten Borgwardt. Unsupervised Manifold Alignment with Joint Multidimensional Scaling. In *The Eleventh International Conference on Learning Representations*, September 2022.

[36] Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised Alignment of Embeddings with Wasserstein Procrustes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, pages 1880–1890. PMLR, April 2019.

[37] Jing Wang, Xueqing Li, and Jixiang Du. Label Space Embedding of Manifold Alignment for Domain Adaption. *Neural Processing Letters*, 49(1):375–391, February 2019. ISSN 1370-4621. doi: 10.1007/s11063-018-9822-8.

[38] 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):MSB20209620, January 2021. ISSN 1744-4292. doi: 10.15252/msb.20209620.

[39] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[40] Peter Halmos, Julian Gold, Xinhao Liu, and Benjamin Raphael. Hierarchical Refinement: Optimal Transport to Infinity and Beyond. In *Forty-second International Conference on Machine Learning*, June 2025.

[41] Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-Rank Sinkhorn Factorization. In *Proceedings of the 38th International Conference on Machine Learning*, pages 9344–9354. PMLR, July 2021.

[42] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, December 2019. ISSN 1546-1696. doi: 10.1038/s41587-019-0336-3. tex.copyright: 2019 The Author(s), under exclusive licence to Springer Nature America, Inc.

[43] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-Visual Event Localization in Unconstrained Videos. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, volume 11206, pages 252–268. Springer International Publishing, Cham, 2018. ISBN 978-3-030-01215-1 978-3-030-01216-8. doi: 10.1007/978-3-030-01216-8_16. URL `https://link.springer.com/10.1007/978-3-030-01216-8_16`. Series Title: Lecture Notes in Computer Science.

[44] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *2011 IEEE International Conference on Robotics and Automation*, pages 1817–1824, May 2011. doi: 10.1109/ICRA.2011.5980382. URL `https://ieeexplore.ieee.org/document/5980382`. ISSN: 1050-4729.

[45] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):119:1–119:12, July 2016. ISSN 0730-0301. doi: 10.1145/2897824.2925954. URL `https://dl.acm.org/doi/10.1145/2897824.2925954`.

[46] Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra, and Jorge L Reyes-Ortiz. A Public Domain Dataset for Human Activity Recognition Using Smartphones. *Computational Intelligence*, 2013.

[47] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[48] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL `http://arxiv.org/abs/1512.03385`. arXiv:1512.03385 [cs.CV].

[49] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, February 2024. URL `http://arxiv.org/abs/2304.07193`. arXiv:2304.07193 [cs.CV].

[50] Felipe A. Vieira Braga, Gozde Kar, Marijn Berg, Orestes A. Carpaij, Krzysztof Polanski, Lukas M. Simon, Sharon Brouwer, Tomás Gomes, Laura Hesse, Jian Jiang, Eirini S. Fasouli, Mirjana Efremova, Roser Vento-Tormo, Carlos Talavera-López, Marnix R. Jonker, Karen Affleck, Subarna Palit, Paulina M. Strzelecka, Helen V. Firth, Krishnaa T. Mahbubani, Ana Cvejic, Kerstin B. Meyer, Kourosh Saeb-Parsy, Marjan Luinge, Corry-Anke Brandsma, Wim Timens, Ilias Angelidis, Maximilian Strunz, Gerard H. Koppelman, Antoon J. van Oosterhout, Herbert B. Schiller, Fabian J. Theis, Maarten van den Berge, Martijn C. Nawijn, and Sarah A. Teichmann. A cellular census of human lungs identifies novel cell states in health and in asthma. *Nature Medicine*, 25(7):1153–1163, July 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0468-5. tex.copyright: 2019 The Author(s), under exclusive licence to Springer Nature America, Inc.

[51] Malte D. Luecken, M. Büttner, K. Chaichoompu, A. Danese, M. Interlandi, M. F. Mueller, D. C. Strobl, L. Zappia, M. Dugas, M. Colomé-Tatché, and Fabian J. Theis. Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41–50, January 2022. ISSN 1548-7105. doi: 10.1038/s41592-021-01336-8. tex.copyright: 2021 The Author(s).

[52] 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.

[53] Huy Q. Tran, Kar Soon Ang, Martin Chevrier, Xin Zhang, Nicholas Y. S. Lee, Mui-Hong Goh, and Jason Chen. A benchmark of batch-effect correction methods for single-cell RNA-sequencing data. *Genome Biology*, 21(1):12, 2020. doi: 10.1186/s13059-019-1850-9.

[54] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10):733–739, 2010. doi: 10.1038/nrg2825.

[55] Kyle J Travaglini, Ahmad N Nabhan, Lolita Penland, Rahul Sinha, Astrid Gillich, Rene V Sit, Stephen Chang, Stephanie D Conley, Yasuo Mori, Jun Seita, and others. A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature*, 587(7835):619–625, 2020.

[56] Micha Sam Brickman Raredon, Taylor Sterling Adams, Yasir Suhail, Jonas Christian Schupp, Sergio Poli, Nir Neumark, Katherine L. Leiby, Allison Marie Greaney, Yifan Yuan, Corey Horien, George Linderman, Alexander J. Engler, Daniel J. Boffa, Yuval Kluger, Ivan O. Rosas, Andre Levchenko, Naftali Kaminski, and Laura E. Niklason. Single-cell connectomic analysis of adult mammalian lungs. *Science Advances*, 5(12):eaaw3851, December 2019. ISSN 2375-2548. doi: 10.1126/sciadv.aaw3851. URL `https://www.science.org/doi/10.1126/sciadv.aaw3851`.

[57] Katherine C. Goldfarbmuren, Nathan D. Jackson, Satria P. Sajuthi, Nathan Dyjack, Katie S. Li, Cydney L. Rios, Elizabeth G. Plender, Michael T. Montgomery, Jamie L. Everman, Preston E. Bratcher, Eszter K. Vladar, and Max A. Seibold. Dissecting the cellular specificity of smoking effects and reconstructing lineages in the human airway epithelium. *Nature Communications*, 11(1):2485, May 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-16239-z. URL `https://www.nature.com/articles/s41467-020-16239-z`.

[58] Maria C. Basil, Fabian L. Cardenas-Diaz, Jaymin J. Kathiriya, Michael P. Morley, Justine Carl, Alexis N. Brumwell, Jeremy Katzen, Katherine J. Slovik, Apoorva Babu, Su Zhou, Madison M. Kremp, Katherine B. McCauley, Shanru Li, Joseph D. Planer, Shah S. Hussain, Xiaoming Liu, Rebecca Windmueller, Yun Ying, Kathleen M. Stewart, Michelle Oyster, Jason D. Christie, Joshua M. Diamond, John F. Engelhardt, Edward Cantu, Steven M. Rowe, Darrell N. Kotton, Harold A. Chapman, and Edward E. Morrisey. Human distal airways contain a multipotent secretory cell that can regenerate alveoli. *Nature*, 604 (7904):120–126, April 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04552-0. URL `https://www.nature.com/articles/s41586-022-04552-0`.

[59] Gilles Louppe. *Understanding random forests: From theory to practice*. Universite de Liege (Belgium), 2014.

[60] Kawther Hassine, Aiman Erbad, and Ridha Hamila. Important complexity reduction of random forest in multi-classification problem. In *2019 15th international wireless communications & mobile computing conference (IWCMC)*, pages 226–231. IEEE, 2019.

[61] Erik Demaine, Adam Hesterberg, Frederic Koehler, Jayson Lynch, and John Urschel. Multidimensional Scaling: Approximation and Complexity. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2568–2578. PMLR, July 2021.

[62] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[63] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1): 3–42, 2006.

[64] J.J. Rodriguez, L.I. Kuncheva, and C.J. Alonso. Rotation Forest: A New Classifier Ensemble Method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1619–1630, October 2006. ISSN 1939-3539. doi: 10.1109/TPAMI.2006.211. URL https://ieeexplore.ieee.org/document/1677518.

[65] Erwan Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3): 1485–1500, 2016.

[66] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[67] Mikhail Belkin and Partha Niyogi. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation*, 15(6):1373–1396, June 2003. ISSN 0899-7667. doi: 10.1162/089976603321780317. URL https://ieeexplore.ieee.org/abstract/document/6789755.

[68] Leon Danon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005: P09008, 9 2005. ISSN 1742-5468. doi: 10.1088/1742-5468/2005/09/P09008. URL https://iopscience.iop.org/article/10.1088/1742-5468/2005/09/P09008https://iopscience.iop.org/article/10.1088/1742-5468/2005/09/P09008/meta.

[69] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. doi: 10.1080/01621459.1971.10482356. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1971.10482356.

[70] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 11 1987. ISSN 0377-0427. doi: 10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

[71] Pia Rautenstrauch and Uwe Ohler. Shortcomings of silhouette in single-cell integration benchmarking. *Nature Biotechnology 2025*, pages 1–5, 7 2025. ISSN 1546-1696. doi: 10.1038/s41587-025-02743-4. URL https://www.nature.com/articles/s41587-025-02743-4.

[72] Maren Büttner, Zhichao Miao, F. Alexander Wolf, Sarah A. Teichmann, and Fabian J. Theis. A test metric for assessing single-cell rna-seq batch correction. *Nature methods*, 16:43–49, 1 2019. ISSN 1548-7105. doi: 10.1038/S41592-018-0254-1. URL https://pubmed.ncbi.nlm.nih.gov/30573817/.

A Justification of the semi-supervised RF-GAP target surrogates

We justify the surrogate target-side multiplicities used in the semi-supervised RF-GAP construction. Since unlabeled target points are not included in the bootstrap samples used to train the forest, their in-bag multiplicities are unobserved and must be replaced by deterministic surrogates.

We work under standard assumptions for Random Forest training: each tree is grown independently using bootstrap sampling with replacement from the labeled subset, and each bootstrap sample has the same size as the labeled dataset.

Lemma 1 (Optimal constant surrogate for target-side multiplicities)**.** *Fix a domain $\mathcal{X}$ and a tree $t \in \mathcal{T}$. Let $\mathcal{I}^l$ denote the labeled index set, and let $c_k(t)$ be the bootstrap multiplicity of $\mathbf{x}_k$ in tree t. Consider the problem of approximating the labeled multiplicities by a constant $c \in \mathbb{R}$ in the least-squares sense:*

$$c(t) = \arg\min_{c \in \mathbb{R}} \sum_{k \in \mathcal{I}^l} \big(c_k(t) - c\big)^2.$$

Then the unique minimizer satisfies $c(t) = 1$.

Proof. The objective is strictly convex in c, and its minimizer is given by the empirical mean:

$$c(t) = \frac{1}{|\mathcal{I}^l|} \sum_{k \in \mathcal{I}^l} c_k(t).$$

Under bootstrap sampling with replacement, the total multiplicity equals the sample size, yielding

$$\sum_{k \in \mathcal{I}^l} c_k(t) = |\mathcal{I}^l| \qquad \Longrightarrow \qquad c(t) = 1.$$

□

This result shows that the unit weight is the exact empirical L^2-optimal constant surrogate for missing target-side multiplicities, which justifies its use in Eq. 1 for unlabeled targets.

B Extension to imbalanced domain sizes and partial matching

We focused on the balanced setting, assuming full overlap between the two domains. In practice, however, many applications involve imbalanced domain sizes ($n \neq m$). Several strategies can be employed to extend our framework to this more general setting.

Recall that in the balanced case, correspondence matching is formulated using HiRef, which seeks a Monge map between domains and therefore enforces a bijection (Section 3.3). In the imbalanced setting, this naturally generalizes to seeking a *surjection* from the larger domain onto the smaller one. Without loss of generality, assume $n > m$, and that the semantic structures present in domain B are contained within domain A. To satisfy the balanced assignment constraint required by HiRef, we first augment the smaller domain by duplicating samples from domain B in the semantic space until both domains contain the same number of points. HiRef is then applied to these augmented domains, producing an intermediate balanced coupling corresponding to a bijective assignment between samples from domain A and the augmented samples from domain B.

Since duplicated samples in the augmented domain correspond to the same original observation in B, the resulting coupling can subsequently be condensed back to the original domain sizes, yielding a sparse coupling matrix $T \in \mathbb{R}^{n \times m}$. By construction, each sample from domain B remains matched to exactly one sample from domain A, while multiple samples from A may now map to the same sample in B. The final transport therefore defines a surjective assignment from the larger domain onto the smaller one, while preserving the hard one-to-one structure imposed by HiRef during the intermediate balanced optimization. This is the strategy adopted in our single-cell alignment experiments involving imbalanced real batches.

In practice, the assumption of full overlap between domains may not hold. For example, the label spaces may only partially overlap, leading to domain-specific structures that should not be matched across domains. In this setting, we first subset each domain to retain only samples that are plausibly

matchable to points in the other domain, and then perform the surjective matching procedure on these subsets while leaving domain-specific points unmatched in the resulting transport plan.

When labels are fully available, the shared subset can be directly inferred from the intersection of the label spaces across domains. When labels are only partially observed, and the label spaces do not fully overlap, stronger assumptions about the missing-label mechanism become necessary. In particular, one must decide whether unlabeled points are expected to belong to the shared label space or instead correspond to domain-specific structures. In the former case, unlabeled samples should be included in the OT problem; in the latter, they should be excluded.

To address this ambiguity, we implemented an automatic strategy based on the underlying forest classifier. Unlabeled samples are first assigned pseudo-labels using the forest out-of-bag (OOB) predictions. If the predicted label belongs to the shared label space, the sample is included in the OT matching; otherwise, it is excluded. While this heuristic provided satisfactory empirical behavior in our single-cell alignment tasks with label masking and potential batch-specific cell types (Section 4.2), we leave a more principled treatment of this problem for future work.

C Equivalence between Euclidean and cosine costs under quadratic normalization

We justify the ℓ_2 normalization in Eq. 2 by showing that squared Euclidean distances between normalized semantic vectors are equivalent to cosine distances computed on the original (unnormalized) semantic representations. This ensures compatibility with Euclidean OT solvers such as HiRef, while preserving the cosine-based geometry commonly used in prior work [5, 11].

Lemma 2. *Let $S_*(i,:), S_*(j,:) \in \mathbb{R}^C$ be two nonzero semantic vectors, and define their ℓ_2-normalized versions*

$$\tilde{S}_*(i,:) = \frac{S_*(i,:)}{\|S_*(i,:)\|_2}, \qquad \tilde{S}_*(j,:) = \frac{S_*(j,:)}{\|S_*(j,:)\|_2}.$$

Then,

$$\|\tilde{S}_*(i,:) - \tilde{S}_*(j,:)\|_2^2 = 2\left(1 - \frac{\langle S_*(i,:), S_*(j,:)\rangle}{\|S_*(i,:)\|_2 \, \|S_*(j,:)\|_2}\right).$$

In particular, the squared Euclidean distance between normalized semantic vectors is proportional to the cosine distance between the original semantic vectors.

Proof. By construction,

$$\|\tilde{S}_*(i,:)\|_2 = \|\tilde{S}_*(j,:)\|_2 = 1.$$

Expanding the squared Euclidean distance yields

$$\|\tilde{S}_*(i,:) - \tilde{S}_*(j,:)\|_2^2 = \|\tilde{S}_*(i,:)\|_2^2 + \|\tilde{S}_*(j,:)\|_2^2 - 2\langle \tilde{S}_*(i,:), \tilde{S}_*(j,:)\rangle = 2 - 2\langle \tilde{S}_*(i,:), \tilde{S}_*(j,:)\rangle.$$

Substituting the definition of the normalized inner product gives

$$\langle \tilde{S}_*(i,:), \tilde{S}_*(j,:)\rangle = \frac{\langle S_*(i,:), S_*(j,:)\rangle}{\|S_*(i,:)\|_2 \, \|S_*(j,:)\|_2},$$

which yields the result. □

D Extended complexity analysis

Without loss of generality, assume balanced domains with $N = n = m$ samples each. FoSTA consists of three main stages: (i) computing label-informed RF-GAP affinities within each domain, (ii) estimating a cross-domain transport plan with HiRef in semantic space, and (iii) producing a joint low-dimensional embedding with Landmark PHATE.

Computing RF-GAP within a domain has time complexity $\mathcal{O}(T\,N \log N \cdot p \;+\; TN(h+\ell))$, where T is the number of trees, p is the number of candidate features evaluated per split, $h \approx \log N$ is the average tree height, and ℓ is the average number of same-leaf interactions per sample and per tree [59, 60, 19]. The associated memory footprint scales as $\mathcal{O}(NT(1+\ell))$ for storing the sparse

leaf-based interactions used to assemble affinities [19]. HiRef computes an approximately bijective Monge map in log-linear time and linear space in N [40]. Finally, Landmark PHATE with M landmarks builds landmark relationships at cost $\mathcal{O}(2NM)$ and stores them in $\mathcal{O}(2NM)$ memory, followed by diffusion and embedding on the landmark set, which costs $\mathcal{O}(M^3)$ [42, 61].

In general, T, p, and ℓ are treated as small constants relative to N, and $M \ll N$ is set to $M = 2000$ by default [42]. Under these assumptions, FoSTA has effectively near-linear runtime and memory scaling in the number of samples. In contrast, MALI, Pamona and related OT-based alignment methods rely on full DPT and/or explicit dense cross-domain cost matrices, leading to at least quadratic time and space complexity in N.

E Empirical scaling validation

To complement the theoretical analysis in Appendix D, we include an empirical scalability study on RGB-D (DINOv2-B features) and Sketchy (ResNet18 features). See Appendix G.1 for dataset details. Table E reports runtime and peak memory for computing the joint embedding with 6k, 12k, and 24k samples.

FoSTA exhibits an approximately linear increase in both runtime and memory as the dataset size grows. MALI is faster at smaller sample sizes due to FoSTA's additional computational overhead, but its runtime grows substantially faster as the number of samples increases. The difference is more pronounced in memory usage: at 24k samples, MALI requires approximately 29 GB on both datasets, compared with 5.6 GB and 4.1 GB for FoSTA on RGB-D and Sketchy, respectively. Overall, these empirical results are consistent with the theoretical complexity analysis in Appendix D, showing that FoSTA maintains more favorable scaling as the dataset size increases.

Table 2: **Runtime and peak-memory scaling across datasets and methods.** Values are means $\pm$ sample standard deviations across successful seed repetitions. Columns are grouped by method, with subcolumns giving the total number of samples across both domains. Runtime is in seconds and memory is in GB (1,024 MB per GB).

		FoSTA			MALI		
Metric	Dataset	6,000	12,000	24,000	6,000	12,000	24,000
Runtime (s)	RGB-D	26.24$\pm$0.20	50.76$\pm$1.24	100.65$\pm$0.46	5.51$\pm$0.13	14.99$\pm$0.22	79.99$\pm$11.45
	Sketchy	39.43$\pm$1.03	74.10$\pm$1.45	149.22$\pm$2.74	5.77$\pm$0.08	15.99$\pm$0.26	77.36$\pm$1.37
Memory (GB)	RGB-D	2.03$\pm$0.22	3.18$\pm$0.17	5.55$\pm$0.06	3.36$\pm$0.21	7.91$\pm$0.01	29.30$\pm$0.20
	Sketchy	1.62$\pm$0.17	2.45$\pm$0.21	4.07$\pm$0.18	3.19$\pm$0.07	7.85$\pm$0.23	29.01$\pm$0.18

F Implementation and hyperparameter settings

All experiments were conducted on a MacBook Pro equipped with an Apple M4 Pro processor (12 CPU cores: 8 performance and 4 efficiency cores) and 24 GB of unified memory, running macOS 26.6.2 (Darwin 25.6.0).

Unless otherwise specified, we use the default hyperparameters from the official implementations of all methods. For FoSTA and MALI, we use a PHATE diffusion scale of $t = 2$ throughout all experiments. For Pamona, we use the label-informed OT correction with $\gamma = 0.5$ [8] throughout all experiments.

For the real-world multimodal experiments in Section 4.1 only, we additionally relax Pamona's convergence parameters to $\epsilon = 0.1$, `max_iter`$= 100$, and `tol`$= 10^{-5}$ to obtain reasonable runtimes. All other Pamona experiments use its default convergence parameters.

FoSTA uses the standard implementations of random forests from `scikit-learn` [62], HiRef [40], and Landmark PHATE [42].

G Real-world multimodal data integration: experimental details and extended results

G.1 Datasets and experimental protocol

We evaluate FoSTA on four real-world multimodal datasets in which the two domains describe the same underlying objects or events through different modalities and share a common class-label space. Ground-truth sample correspondences are retained exclusively for evaluation and are not provided to the alignment methods.

- **Audio-Visual Event (AVE [43]).** The dataset contains 4,143 paired audio-video observations spanning 28 event classes, represented by 128 audio and 512 visual features. Correspondence is defined by the underlying event instance. We use the provided features and predefined train/test split: training labels are available during alignment, while test labels are masked and retained for label transfer evaluation. No additional feature normalization or dimensionality reduction is applied.
- **RGB-D Object [44].** Paired RGB and depth observations represent physical objects from 51 classes. From the 207,920 available observations, we use at most 15,000 paired samples for each experiment. Following Duque et al. [11], each modality is encoded using an ImageNet-pretrained ResNet18 [48], with global average pooling of the final convolutional representation yielding 512 features per domain. We construct a class-stratified split with 50% of the labels retained and 50% masked, using the same split in both domains.
- **Sketchy [45].** Photographs and human-drawn sketches span 125 object classes, with multiple sketches potentially corresponding to each photograph. For each seed, we randomly select one corresponding sketch per photograph, yielding 12,500 paired observations. Following Duque et al. [11], both modalities are encoded using an ImageNet-pretrained ResNet18 [48], yielding 512 features per domain. Photographs are treated as the source domain, with all labels retained, while 50% of the sketch target labels are masked within each class.
- **Human Activity Recognition (HAR [46]).** The dataset contains 10,299 paired observations spanning six activity classes. We form the two domains from accelerometer and gyroscope measurements, each represented by 384 features, with paired observations corresponding to the same activity instance. The predefined train/test split determines label availability: training labels are provided during alignment and test labels are masked. Features are standardized independently within each modality to zero mean and unit variance.
- **Alternative image representations.** For RGB-D and Sketchy, we additionally replace ResNet18 with DINOv2-Base [49], representing each image by its 768-dimensional CLS token. All sample selection, label masking, and evaluation procedures remain unchanged, isolating the effect of the input image representation.

To evaluate the methods under partial label supervision, HAR and AVE use their predefined train/test splits, while RGB-D uses a stratified 50% labeled split shared across both domains. For Sketchy, all photo labels are retained and 50% of the sketch labels are masked, following the source-to-target semi-supervised setting of Duque et al. [11]. Models are fit using both labeled and masked samples, with only the available labels used for supervision, and label transfer is evaluated on the held-out labels. Since Sketchy contains multiple sketches per photo, we randomly select one corresponding sketch per photo in each seed to obtain one-to-one ground-truth correspondences for evaluation. Experiments are repeated over five random seeds.

G.2 Evaluation metrics with ground-truth correspondences

Label transfer accuracy is measured by predicting masked labels using a k-NN classifier trained on the opposite domain embeddings [31, 7]. When labels are masked in both domains, label transfer is evaluated in both directions and the two accuracies are averaged; when a natural source and target are defined, as for Sketchy (Appendix G.1) and our UCI experiments (Appendix H), we evaluate source-to-target transfer only. Domain mixing is quantified by an alignment score (AS [47]) defined as $2(1 - \bar{x}/k)$, where $\bar{x}$ is the average number of k-NN belonging to the same domain as the query sample. An AS near 2 corresponds to $\bar{x} \approx 0$, meaning that a point's neighbors come almost entirely

from the other domain; AS near 1 corresponds to $\bar{x} \approx k/2$, meaning balanced local mixing across domains; and AS near 0 corresponds to $\bar{x} \approx k$, meaning complete separation. Correspondence recovery is assessed using the Fraction of Samples Closer Than the True Match (FOSCTTM [30]), defined as the fraction of cross-domain samples that are closer than the true correspondence (lower is better).

H Synthetic multimodal data integration benchmarks and ablation experiments

H.1 Datasets and experimental protocol

We use 15 supervised learning datasets from the UCI Machine Learning Repository to evaluate synthetic multimodal data integration performance in Sections H.2, H.3, and H.4. Following the experimental protocol of Rhodes and Rustad [15], we select the Balance Scale, Breast Cancer, CRX, Diabetes, E.coli (restricted to five majority classes), Flare1, Glass, Heart Disease, Heart Failure, Hepatitis, Ionosphere, Iris, Parkinson's, Seeds, and Tic-Tac-Toe datasets. Observations with missing values are removed prior to analysis. Table 3 summarizes each dataset, including the number of observations, features, and classes.

To construct two domains from each dataset, we consider six controlled perturbations [15]. The *random* split randomly assigns features to either domain. The *importance* split assigns the most task-relevant features, as determined by a random forest model, to the first domain, and the least important features to the second domain. The *alternating importance* split assigns features to either domain by alternating by feature importance, creating balanced domains. In the *add noise* split, both domains contain all of the features, but the second additionally contains Gaussian noise features ($\sigma = 1$), with a signal-to-noise ratio of 1 to 10. The *distort* split creates a second domain by adding Gaussian noise ($\sigma = 0.5$) to the first, while the *rotate* split creates a second domain by a random rotation of the first.

Unless otherwise specified, we mask 50% of the target-domain labels to evaluate the methods in a semi-supervised setting. We use five different random seeds to generate the domains and report the averages. As in Section 4.1, we evaluate the methods using label-transfer accuracy, domain mixing (AS), and correspondence recovery (FOSCTTM). Section H.2 compares the methods across the six controlled perturbations, while Sections H.3 and H.4 respectively evaluate the robustness of FoSTA to secondary design choices and varying levels of label availability.

Table 3: **Summary of the UCI datasets used to evaluate synthetic multimodal data integration performance.** Observations with missing values were removed prior to analysis.

	Balance Scale	*Breast Cancer*	*CRX*	*Diabetes*	*E. coli*	*Flare1*	*Glass*	*Heart Disease*	*Heart Failure*	*Hepatitis*	*Ionosphere*	*Iris*	*Parkinson's*	*Seeds*	*Tic-Tac-Toe*
# Obs.	625	699	664	678	336	323	214	303	299	138	351	150	195	199	958
# Features	4	16	14	8	8	10	9	13	10	15	34	4	22	7	9
# Classes	3	2	2	2	5	3	6	5	2	2	2	3	2	3	2

H.2 Performance under controlled perturbations

In Table 4, FoSTA achieves the best or second-best accuracy in five of the six splits (4th in rotate) and the best FOSCTTM in all but two (3rd in distort and rotate). FoSTA also ranks either second or third in AS. The rotate split remains the most challenging regime for FoSTA. This is not surprising: standard tree ensembles rely on axis-aligned splits, so a global rotation can substantially alter the induced forest geometry even when the underlying class semantics remain unchanged. These results indicate strong correspondence recovery and preservation of task-relevant semantic structure across different controlled perturbations.

We emphasize label-transfer accuracy and correspondence recovery over AS, as these metrics are directly grounded in available ground-truth information. Label transfer evaluates whether semantic

Table 4: **Aggregated model performance over UCI datasets and seeds under selected distortion splits.** Each cell reports mean ± the average within-dataset standard deviation across five seeds. Results are shown for label transfer accuracy, alignment score (AS), and correspondence recovery measured by FOSCTTM. Higher is better for accuracy and AS, while lower is better for FOSCTTM. The top three results for each metric are highlighted in gold (1st), silver (2nd), and bronze (3rd). The final three rows report average scores across the displayed splits, with standard deviations also averaged across splits.

Split	Metric	**FoSTA**	KEMAlin	KEMArbf	MALI	Pamona
Noise	Accuracy↑	**0.788**±0.02	**0.690**±0.03	0.671±0.04	**0.691**±0.04	0.615±0.06
	AS↑	**0.294**±0.02	0.287±0.04	0.290±0.03	**0.883**±0.02	**0.936**±0.02
	FOSCTTM↓	**0.251**±0.02	0.339±0.01	**0.335**±0.01	**0.329**±0.02	0.362±0.03
Alt. Imp.	Accuracy↑	**0.767**±0.03	0.708±0.05	**0.751**±0.03	**0.718**±0.04	0.703±0.05
	AS↑	**0.732**±0.04	0.274±0.08	0.564±0.06	**0.950**±0.03	**0.567**±0.05
	FOSCTTM↓	**0.313**±0.02	0.379±0.02	**0.334**±0.01	0.368±0.03	**0.362**±0.03
Distort	Accuracy↑	**0.773**±0.03	**0.766**±0.02	0.758±0.03	0.761±0.03	**0.786**±0.03
	AS↑	**0.643**±0.04	0.409±0.04	0.553±0.04	**1.001**±0.02	**0.965**±0.02
	FOSCTTM↓	**0.238**±0.02	0.288±0.02	0.303±0.01	**0.211**±0.02	**0.156**±0.05
Importance	Accuracy↑	**0.720**±0.03	**0.698**±0.03	**0.722**±0.03	0.684±0.04	0.664±0.04
	AS↑	**0.615**±0.04	0.263±0.05	0.477±0.03	**0.857**±0.02	**0.533**±0.03
	FOSCTTM↓	**0.329**±0.01	0.387±0.01	**0.345**±0.01	0.378±0.03	**0.370**±0.02
Random	Accuracy↑	**0.775**±0.04	0.724±0.05	**0.756**±0.05	**0.727**±0.04	0.694±0.05
	AS↑	**0.707**±0.06	0.287±0.08	0.555±0.08	**0.946**±0.03	**0.568**±0.07
	FOSCTTM↓	**0.316**±0.02	0.384±0.02	**0.337**±0.02	0.374±0.03	**0.372**±0.03
Rotate	Accuracy↑	0.807±0.03	0.794±0.02	**0.808**±0.02	**0.819**±0.02	**0.832**±0.03
	AS↑	**0.716**±0.05	0.471±0.07	0.600±0.05	**1.042**±0.02	**1.018**±0.01
	FOSCTTM↓	**0.218**±0.02	0.242±0.03	0.266±0.02	**0.109**±0.04	**0.075**±0.02
Average score	Accuracy↑	**0.772**±0.03	0.730±0.03	**0.744**±0.03	**0.733**±0.03	0.716±0.04
	AS↑	**0.618**±0.04	0.332±0.06	0.507±0.05	**0.947**±0.02	**0.765**±0.03
	FOSCTTM↓	**0.278**±0.02	0.337±0.02	0.320±0.01	**0.295**±0.03	**0.283**±0.03

information is preserved across domains, while FOSCTTM explicitly measures the recovery of true cross-domain correspondences. In contrast, AS is a heuristic that assesses domain mixing under the assumption that local neighborhoods should contain uniformly distributed samples from each domain. As a result, a method could achieve a high AS by indiscriminately mixing domains—potentially even collapsing structure—without accurately aligning shared semantic manifolds. Thus, a moderate AS does not contradict strong performance in label transfer and correspondence recovery, which provide a more direct assessment of alignment quality in our setting.

H.3 Robustness to secondary design choices

This section presents an ablation study of FoSTA across secondary architectural choices in the pipeline, following the upstream-to-downstream construction described in the Methods. Table 5 summarizes the results under the same experimental protocol as in Section H.2 on the UCI benchmark. For each variant, we report rankings across splits and metrics relative to the baseline competitors in Table 4.

- **Upstream geometry (Random Forests vs ExtraTrees vs Rotation Forests).** Replacing Random Forests with Extremely Randomized Trees (ExtraTrees [63]) consistently improves performance, particularly in AS and FOSCTTM. We hypothesize that, on relatively simple prediction tasks, the additional randomness introduced by ExtraTrees leads to coarser yet informative partitions of the feature space. This produces semantic representations that better localize samples in the supervised manifold, whereas Random Forests may be overly

Table 5: **FoSTA ablation performance over UCI datasets and seeds under the selected distortion splits.** Each cell reports mean ± the average within-dataset standard deviation across five seeds. Results are shown for label transfer accuracy (Acc), alignment score (AS), and correspondence recovery measured by FOSCTTM (FOS). Higher is better for accuracy and AS, while lower is better for FOSCTTM. The top three results for each metric are highlighted in gold (1st), silver (2nd), and bronze (3rd). Each variant is ranked independently against non-FoSTA baselines in Table 4. The final three rows report average scores across the displayed splits, with standard deviations also averaged across splits.

Split	Metric	**Default**	RotF	ET	KeRF	DPT	Dense OT	No Prop.	Auto. t	UMAP	LE
Noise	Acc↑	**0.788**±0.02	**0.799**±0.02	**0.788**±0.03	**0.768**±0.03	**0.783**±0.03	**0.784**±0.02	0.629±0.08	**0.770**±0.03	**0.787**±0.03	**0.769**±0.02
	AS↑	**0.294**±0.02	**0.462**±0.04	0.278±0.03	**0.312**±0.02	**0.315**±0.03	0.230±0.02	0.000±0.00	**0.448**±0.02	**0.973**±0.01	**0.546**±0.03
	FOS↓	**0.251**±0.02	**0.230**±0.02	**0.247**±0.02	**0.253**±0.02	**0.244**±0.02	**0.242**±0.02	**0.333**±0.01	**0.251**±0.01	**0.302**±0.01	**0.270**±0.02
Alt. Imp.	Acc↑	**0.767**±0.03	**0.766**±0.03	**0.763**±0.02	**0.760**±0.03	**0.755**±0.03	**0.769**±0.03	**0.735**±0.04	**0.769**±0.02	**0.769**±0.02	**0.760**±0.02
	AS↑	**0.732**±0.04	**0.693**±0.05	**0.820**±0.04	**0.717**±0.05	**0.746**±0.05	**0.732**±0.05	0.154±0.04	**0.767**±0.05	**0.945**±0.02	**0.775**±0.05
	FOS↓	**0.313**±0.02	**0.307**±0.01	**0.309**±0.02	**0.317**±0.02	**0.320**±0.02	**0.307**±0.02	**0.335**±0.01	**0.308**±0.02	**0.321**±0.02	**0.332**±0.02
Distort	Acc↑	**0.773**±0.03	**0.776**±0.03	**0.779**±0.03	**0.763**±0.03	**0.767**±0.03	**0.770**±0.03	0.697±0.05	**0.767**±0.03	**0.774**±0.03	**0.764**±0.03
	AS↑	**0.643**±0.04	**0.629**±0.03	**0.717**±0.04	**0.616**±0.03	**0.647**±0.03	**0.657**±0.04	0.003±0.00	**0.746**±0.03	**0.973**±0.02	**0.783**±0.03
	FOS↓	**0.238**±0.02	**0.226**±0.02	**0.225**±0.02	**0.235**±0.02	**0.225**±0.02	**0.232**±0.02	0.304±0.01	**0.231**±0.01	**0.271**±0.01	**0.258**±0.02
Importance	Acc↑	**0.720**±0.03	**0.726**±0.03	**0.724**±0.03	**0.722**±0.03	0.679±0.05	**0.724**±0.04	0.661±0.06	**0.723**±0.03	**0.720**±0.04	**0.717**±0.03
	AS↑	**0.615**±0.04	**0.598**±0.04	**0.694**±0.04	**0.597**±0.04	**0.641**±0.04	**0.605**±0.05	0.118±0.03	**0.668**±0.04	**0.912**±0.02	**0.702**±0.03
	FOS↓	**0.329**±0.01	**0.321**±0.01	**0.327**±0.01	**0.330**±0.02	**0.352**±0.02	**0.325**±0.01	**0.353**±0.01	**0.324**±0.01	**0.334**±0.01	**0.344**±0.01
Random	Acc↑	**0.775**±0.04	**0.780**±0.04	**0.772**±0.05	**0.772**±0.04	**0.739**±0.06	**0.776**±0.04	**0.740**±0.07	**0.774**±0.04	**0.769**±0.05	**0.771**±0.05
	AS↑	**0.707**±0.06	**0.661**±0.07	**0.787**±0.04	**0.686**±0.06	**0.707**±0.06	**0.706**±0.06	0.129±0.03	**0.760**±0.06	**0.942**±0.02	**0.773**±0.05
	FOS↓	**0.316**±0.02	**0.309**±0.02	**0.313**±0.02	**0.313**±0.02	**0.335**±0.04	**0.311**±0.02	**0.336**±0.02	**0.309**±0.02	**0.322**±0.02	**0.335**±0.02
Rotate	Acc↑	0.807±0.03	**0.820**±0.02	**0.813**±0.02	**0.808**±0.02	0.804±0.03	0.807±0.02	0.738±0.05	0.805±0.03	**0.808**±0.02	0.804±0.03
	AS↑	**0.716**±0.05	**0.656**±0.04	**0.826**±0.04	**0.671**±0.04	**0.707**±0.04	**0.731**±0.04	0.010±0.01	**0.798**±0.05	**0.962**±0.02	**0.822**±0.04
	FOS↓	**0.218**±0.02	**0.189**±0.01	**0.196**±0.02	**0.221**±0.02	**0.213**±0.03	**0.208**±0.02	0.289±0.02	**0.207**±0.01	0.246±0.02	**0.240**±0.02
Average score	Acc↑	**0.772**±0.03	**0.778**±0.03	**0.773**±0.03	**0.766**±0.03	**0.755**±0.04	**0.772**±0.03	0.700±0.06	**0.768**±0.03	**0.771**±0.03	**0.764**±0.03
	AS↑	**0.618**±0.04	**0.616**±0.04	**0.687**±0.04	**0.600**±0.04	**0.627**±0.04	**0.610**±0.04	0.069±0.02	**0.698**±0.04	**0.951**±0.02	**0.733**±0.04
	FOS↓	**0.278**±0.02	**0.264**±0.01	**0.270**±0.02	**0.278**±0.02	**0.282**±0.02	**0.271**±0.02	0.325±0.01	**0.272**±0.01	**0.299**±0.02	**0.296**±0.02

confident despite the presence of bootstrap variability. Rotation Forests (RotF [64]) yield similar overall performance to the default Random Forest geometry, with slightly improved accuracy and FOSCTTM on average. The improvement is most apparent under the rotation distortion, supporting the interpretation that part of FoSTA's sensitivity in this setting comes from the axis-aligned splits of standard Random Forests rather than from the alignment procedure itself.

- **Affinity construction (semi-supervised RF-GAP vs KeRF).** KeRF affinities [65] can be interpreted as a symmetric variant of RF-GAP, where pairwise similarities are still defined through leaf collisions, but without distinguishing between in-bag and out-of-bag interactions, and with leaf mass normalization computed over the full training set. KeRF yields slightly lower accuracy and AS than the default setting, while achieving comparable FOSCTTM. Overall, these results support semi-supervised RF-GAP as the default affinity construction, while showing that FoSTA remains effective with alternative forest-based affinities.
- **Diffusion coarse-graining (DPT vs no DPT).** Applying Diffusion Pseudotime (DPT) to the RF-GAP affinities before semantic matching provides little improvement in AS while degrading accuracy and FOSCTTM. This additional diffusion step is therefore both computationally unnecessary and detrimental to correspondence recovery. More importantly, these results suggest that the forest affinities already provide the graded, local-to-global class-aware structure needed for semantic optimal transport, without requiring additional diffusion-based preprocessing.
- **Alignment stage (HiRef vs dense OT).** HiRef and exact balanced Wasserstein OT with uniform masses yield comparable performance. This indicates that HiRef offers a favorable trade-off between computational efficiency and alignment quality, integrating well within the FoSTA pipeline while preserving near-linear scalability.
- **Affinity propagation (with vs without propagation).** Removing affinity propagation substantially degrades performance, notably reducing AS from 0.618 to 0.069 on average, while also worsening accuracy and FOSCTTM. This confirms that the semantic correspondences

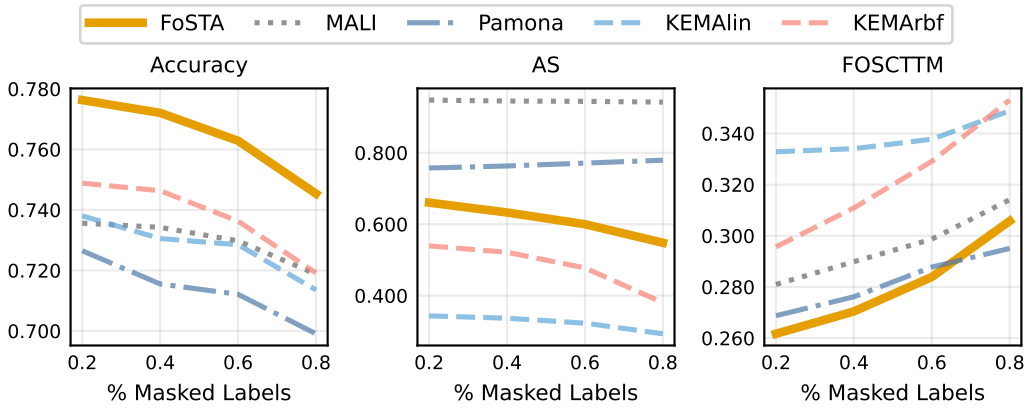

Figure 3: Aggregated performance across 15 UCI datasets as a function of the percentage of available labels in the target domain.

identified by optimal transport alone are insufficient: propagating them through the forest affinities is important to extend the class-level matching to the full sample geometry and obtain a coherent cross-domain alignment.

- **Downstream embedding (default PHATE ($t = 2$) vs automatic t selection vs UMAP [66] vs Laplacian Eigenmaps (LE [67]).** Automatic entropy-based scale selection in PHATE [42] improves AS and slightly improves FOSCTTM, at the expense of some label-transfer accuracy. This is consistent with stronger diffusion promoting mixing across the cross-domain affinity blocks, while also smoothing local class structure. UMAP substantially improves AS and ranks first overall on this metric, but yields worse FOSCTTM, highlighting a trade-off between global domain mixing and accurate correspondence recovery, consistent with Section 4.1. LE also improves AS over the default PHATE embedding, but decreases both accuracy and FOSCTTM.

Overall, FoSTA remains competitive across forest, transport, and embedding variants, although individual choices such as UMAP and LE shift the relative performance across metrics. Affinity propagation is a notable exception, with its removal substantially degrading alignment and confirming its importance in transferring semantic correspondences through the forest geometry. Together, these results indicate that FoSTA is generally robust to secondary design choices, while supporting that its main advantage over Euclidean-based approaches such as MALI stems from the upstream forest-guided geometry and its propagation across domains.

H.4 Robustness to partial label availability

We evaluate the robustness of FoSTA to varying levels of label availability in the target ("weak") domain. Following the protocol of Section H.2, we vary the proportion of masked target labels from 20% to 80% in increments of 20%, rather than fixing it at 50% as in our previous UCI experiments.

The results are reported in Fig. 3. FoSTA consistently ranks among the top methods in Accuracy and FOSCTTM across masking levels, falling to second in FOSCTTM only under the lowest label-availability regimes, while maintaining the third-best AS throughout. Performance degrades gradually as fewer labels become available, without a marked drop relative to competing methods. Overall, these results show that FoSTA remains competitive even under limited label supervision.

I Batch effect removal in single-cell genomics: experimental details and extended results

I.1 Datasets and experimental protocol

We use the lung atlas dataset from Vieira Braga et al. [50], following the benchmark setting of Luecken et al. [51]. The 16 batches are organized into three families: A (A1–A6), B (B1–B4), and 1–6. We consider all pairwise combinations within each family, resulting in 36 real batch-alignment tasks. Each pair is treated as two separate domains and aligned independently. Paired batches do not necessarily contain the same cell types, allowing evaluation under partial overlap of the shared semantic space.

The data are scran pool-normalized, and we retain 2,000 highly variable genes selected across each pair using the `cell_ranger` flavor. To evaluate the methods under partial label supervision, we independently mask 20% of the cell-type labels within each batch during alignment, while ensuring that at least 10 cells of every masked cell type remain unmasked. Masking is performed independently for each batch, pair, and random seed.

Experiments are repeated over five random seeds. As our evaluation focuses on joint visualization, all methods produce two-dimensional embeddings. Biological conservation and batch correction are computed independently for each batch pair and seed before aggregation across the 36 batch pairs and five seeds.

I.2 Evaluation metrics

The implementation of all metrics for batch correction and biological conservation is taken from the scib-metrics package [51]. All metrics are normalized and transformed appropriately so that the result lies between 0 and 1, with higher scores corresponding to better performance.

For all definitions, we assume the following notation. Denote the dataset by $\mathcal{D} = \{x_1, ..., x_n\}$, with $x_i \in \mathbb{R}^d$ for all i. Let $X \in \mathbb{R}^{n \times d}$ denote the data matrix, with samples in the rows and n denoting the number of samples. Let $\mathcal{L} = \{1, ..., n_l\}$ denote a set of n_l (encoded) labels, which could be the cell type labels or the batch label, depending on the case, and let $L \in \mathcal{L}^n$ denote the vector of labels for each sample.

I.2.1 Biological conservation metrics

Normalized mutual information (NMI [68]) and Adjusted Rand index (ARI [69]) measure the preservation of clustering, i.e., how well a clustering of the integrated data overlaps with the cell type labels.

NMI is computed as follows. Let $\mathcal{C} = \{1, ..., n_c\}$ denote a set of n_c clusters from k-means in the embedding under evaluation, and let $C \in \mathcal{C}^n$ denote the vector of cluster assignments for each sample. Let N be the confusion matrix with rows corresponding to cell type labels, and columns corresponding to clusters in the embedding under evaluation, with entries n_{lc} equal to the number of samples with cell type label l that appear in cluster c of the embedding. Now let $n_{l\cdot} = \sum_i n_{lj}$ denote the total number of samples in cell type l, $n_{\cdot c} = \sum_i n_{ic}$ denote the total number of samples in cluster c, and $n = \sum_{ij} n_{ij}$ denote the total number of cells.

Then, the mutual information of labels and clusters in the confusion matrix N is given by

$$\text{MI}(L, C) = \sum_{l \in \mathcal{L}} \sum_{c \in \mathcal{C}} n_{lc} \log \frac{n \cdot n_{lc}}{n_{\cdot c} n_{l\cdot}}.$$

Define the entropy of the cell type labels as $H(L) = -\sum_{l \in \mathcal{L}} n_{l\cdot} \log(n_{l\cdot}/n)$, and the entropy of the clusters as $H(C) = -\sum_{c \in \mathcal{C}} n_{\cdot c} \log(n_{\cdot c}/n)$. Then,

$$\text{NMI}(L, C) = \frac{2 \cdot \text{MI}(N)}{H(L) + H(C)}.$$

ARI is another metric to evaluate the alignment of two k-means clusterings, based on comparing the amount of agreement to that of a random clustering. Let

$$n_{\text{agreements}} = \sum_{i,j} \mathbb{I}(L_i = L_j \text{ and } C_i = C_j) + \sum_{i,j} \mathbb{I}(L_i \neq L_j \text{ and } C_i \neq C_j).$$

Define the number of of true positives n_{tp}, false positives n_{fp}, true negatives n_{tn}, and false negatives n_{fn} as

$$n_{\text{tp}} = \sum_{i,j} \mathbb{I}(L_i = L_j \text{ and } C_i = C_j), \qquad n_{\text{fp}} = \sum_{i,j} \mathbb{I}(L_i \neq L_j \text{ and } C_i = C_j),$$

$$n_{\text{tn}} = \sum_{i,j} \mathbb{I}(L_i \neq L_j \text{ and } C_i \neq C_j), \qquad n_{\text{tn}} = \sum_{i,j} \mathbb{I}(L_i = L_j \text{ and } C_i \neq C_j).$$

The ARI is then given by Pedregosa et al. [62]:

$$ARI(L, C) = \frac{2(n_{\text{tp}} n_{\text{tn}} - n_{\text{fn}} n_{\text{fp}})}{(n_{\text{tp}} + n_{\text{fn}}) \cdot (n_{\text{fn}} + n_{\text{tn}}) + (n_{\text{tp}} + n_{\text{fp}}) \cdot (n_{\text{fp}} + n_{\text{tn}})}.$$

Silhouette label evaluates cell type separation in the integrated embedding. It is computed with the Average Silhouette Width (ASW [70]) of cell types, i.e., the average over all samples i of

$$ASW(L) = \frac{1}{n} \sum_{i \in \mathcal{D}} s(i) = \frac{1}{n} \sum_{i \in \mathcal{D}} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where

$$a(i) = \frac{1}{|\{j \in \mathcal{D} | L_j = L_i\}|} \sum_{j \in \mathcal{D} | L_j = L_i} d(i, j)$$

denotes the average distance to cells in the same cluster, and

$$b(i) = \min_{L_k \neq L_i} \frac{1}{|\{j \in \mathcal{D} | L_j = L_k\}|} \sum_{j \in \mathcal{D} | L_j = L_k} d(i, j)$$

denotes the minimum average distance to another cluster. For silhouette label, L is the vector of cell type labels.

cLISI is based on the local Inverse Simpson's Index (LISI), a measure of diversity that captures how many samples must be drawn from a neighborhood of a sample before drawing a duplicate label. For each cell, a probability distribution p_i over its neighborhood is defined based on a Gaussian kernel of the distances to the cell of interest. Then

$$\text{LISI} = \text{median}_{i \in \mathcal{D}} \frac{1}{\sum_{l \in \mathcal{L}} p_i(l)},$$

where $p_i(l) = \sum_j p_i(j) \mathbb{I}(L_j = l)$. The result is transformed so that high scores correspond to low cell type diversity (high purity). cLISI [33] is the LISI where L is the cell types, measuring the level of cell type mixing that occurs in the neighborhoods of all cells.

The **isolated labels** metric evaluates the separation of isolated labels, defined as labels that are present in the fewest batches, from other labels. It is computed as the ASW of the isolated labels versus the other labels.

I.2.2 Batch correction metrics

Batch removal adapted silhouette (**BRAS** [71]) is an adaptation of ASW where L is the batch identity labels that is better suited to evaluate batch mixing and nested batch effects. BRAS is defined like ASW, except that $b(i)$ is defined as the mean distance to cells in other clusters: $b(i) = \frac{1}{|\{j \in \mathcal{D} | L_j \neq L_k\}|} \sum_{j \in \mathcal{D} | L_j \neq L_k} d(i, j)$.

Then,

$$\text{BRAS} = \frac{1}{n_l} \sum_{l \in \mathcal{L}} \left[\frac{1}{|\{i \in \mathcal{D} | L_j = l\}|} \sum_{\{i \in \mathcal{D} | L_j = l\}} (1 - |s(i)|) \right]$$

iLISI [33] is the median of LISI applied to the batch labels (see section I.2.1).

kBET [72] performs statistical tests to determine whether local composition of neighborhoods in terms of batch labels corresponds to the global batch distribution. The output is the average rejection rate of the null hypothesis (complete batch mixing) computed over the neighborhoods of samples selected at random. Here, kBET is computed separately for each cell type, and then averaged over all cell types:

$$\text{kBET}(\mathcal{D}) = \frac{1}{n_l} \sum_{l \in \mathcal{L}} \text{kBET}(l) = \frac{1}{n_l} \sum_{l \in \mathcal{L}} \left(\frac{1}{|\mathcal{Y}|} \sum_{i \in \mathcal{Y}} S_i^l \right),$$

where S_i^l is the rejection rate of the statistical test on the batch labels of a neighborhood centered at i restricted to cells with label l, and $\mathcal{Y} \subset \mathcal{D}$ is a random subset of samples.

Graph connectivity measures, for each label l, the connectivity of a k-NN graph G_l restricted to samples with label l, i.e.,

$$GC = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \frac{|LCC(G_l)|}{n_l},$$

where LCC denotes the largest connected component of a graph and n_l denotes the number of samples with label l.

Principal Component Regression (**PCR** [72]) measures the variance explained by the batch. It is based on Principal component analysis (PCA), which returns n principal components (PC). By fitting a linear regression of the batches on the ith PC, we obtain an R^2 value, denoted R_i^2. The total variance explained by the batch is the weighted average over all PCs of the variance explained by the PC with the R^2 of the batch on this PC:

$$\text{Var} = \sum_i \text{Var}_i \cdot R_i^2,$$

where Var_i denotes the variance of the data matrix X explained by the ith principal component.

I.3 Complete quantitative results and extended visualizations

Table 6 provides the complete quantitative results for all methods and metrics in the batch effect correction task described in Section 4.2.

To complement the qualitative assessment in Section 4.2, Fig. 4 presents an extended comparison of two-dimensional embeddings for the B2/B3 batch pair (seed 56089). In addition to FoSTA, scANVI, and the unintegrated PCA representation shown in Fig. 2, we include scVI, MALI, and KEMArbf. MALI and KEMArbf are included as the next highest-ranking methods after scANVI and scVI in the aggregated quantitative comparison, while the unintegrated representation is retained as a control.

Consistent with the quantitative results, FoSTA achieves good batch mixing while preserving a structured embedding with clearly distinguishable cell types. MALI also achieves strong batch correction and retains meaningful global structure, although cell types are less clearly separated. scANVI and scVI preserve biological groupings but produce more isolated cluster configurations, while KEMArbf strongly mixes the batches at the expense of biological structure. Overall, the visualization illustrates the trade-off between batch correction and biological conservation captured by the quantitative metrics.

I.4 Robustness to partial label availability

In Fig. 5, we repeat the batch alignment experiments from Section 4.2 under reduced label availability, masking 50% of the labels instead of 20%. As expected with less supervision, FoSTA shows a slight decrease in biological conservation. Interestingly, this is accompanied by slightly improved batch correction, resulting in a similar overall balance between the two objectives: FoSTA ranks third in

Table 6: **Biological conservation and batch correction on the lung integration benchmark.** Results are mean ± average within-pair standard deviation across seeds. Standard deviations are computed separately within each selected folder and batch pair, then averaged across pairs; higher is better for every metric. Methods are ordered identically in both sections by average Bio/Batch rank, with alphabetical tie-breaking. Gold, silver, and bronze mark the top three distinct scores per metric; ties share a highlight.

Biological conservation						
Method	Isolated ↑	NMI ↑	ARI ↑	ASW label ↑	cLISI ↑	Bio score ↑
FoSTA	0.490±0.02	**0.628±0.01**	**0.525±0.03**	0.452±0.01	0.989±0.00	**0.617±0.01**
scANVI	**0.560±0.02**	**0.735±0.01**	**0.498±0.04**	**0.495±0.01**	**1.000±0.00**	**0.658±0.01**
scVI	**0.535±0.01**	**0.627±0.01**	**0.387±0.03**	0.433±0.01	**0.994±0.00**	**0.595±0.01**
MALI	0.495±0.02	0.513±0.02	0.312±0.03	0.377±0.01	0.961±0.00	0.532±0.01
KEMArbf	0.464±0.02	0.297±0.06	0.170±0.07	0.430±0.01	0.912±0.02	0.455±0.03
Scanorama	0.473±0.00	0.595±0.00	0.382±0.00	**0.460±0.00**	0.983±0.00	0.579±0.00
LIGER	0.467±0.00	0.470±0.01	0.291±0.01	0.386±0.00	0.949±0.00	0.513±0.00
Pamona	**0.523±0.02**	0.548±0.02	0.286±0.02	0.393±0.01	**0.990±0.00**	0.548±0.01
Unintegrated	0.492±0.00	0.579±0.00	0.365±0.00	**0.454±0.00**	0.984±0.00	0.575±0.00
KEMAlin	0.482±0.02	0.311±0.04	0.180±0.05	0.421±0.01	0.884±0.01	0.456±0.02

Batch correction						
Method	BRAS ↑	iLISI ↑	kBET ↑	Graph ↑	PCR ↑	Batch score ↑
FoSTA	0.292±0.02	0.484±0.03	0.465±0.05	0.716±0.03	**0.966±0.00**	**0.585±0.02**
scANVI	**0.349±0.02**	0.390±0.03	**0.699±0.06**	**0.953±0.02**	0.515±0.17	0.581±0.04
scVI	0.316±0.02	0.444±0.03	0.528±0.05	**0.843±0.02**	0.750±0.09	0.576±0.03
MALI	0.269±0.02	**0.695±0.02**	0.446±0.05	0.710±0.03	**0.930±0.03**	**0.610±0.02**
KEMArbf	**0.345±0.03**	**0.569±0.05**	**0.628±0.07**	0.656±0.04	**0.881±0.04**	**0.616±0.02**
Scanorama	**0.345±0.00**	0.391±0.00	**0.585±0.00**	0.748±0.00	0.283±0.00	0.470±0.00
LIGER	0.301±0.00	**0.656±0.02**	0.482±0.03	0.570±0.01	0.721±0.01	0.546±0.01
Pamona	0.307±0.02	0.283±0.02	0.404±0.06	**0.767±0.02**	0.785±0.10	0.509±0.03
Unintegrated	0.334±0.00	0.335±0.00	0.516±0.00	0.729±0.00	0.304±0.00	0.444±0.00
KEMAlin	0.322±0.02	0.468±0.05	0.437±0.07	0.603±0.04	0.719±0.07	0.510±0.03

biological conservation and second in batch correction, compared with second and third, respectively, under 20% masking. Its average rank therefore remains unchanged at 2.5, tied for best overall with scVI and scANVI. These results suggest that FoSTA remains robust even when substantially fewer labels are available.

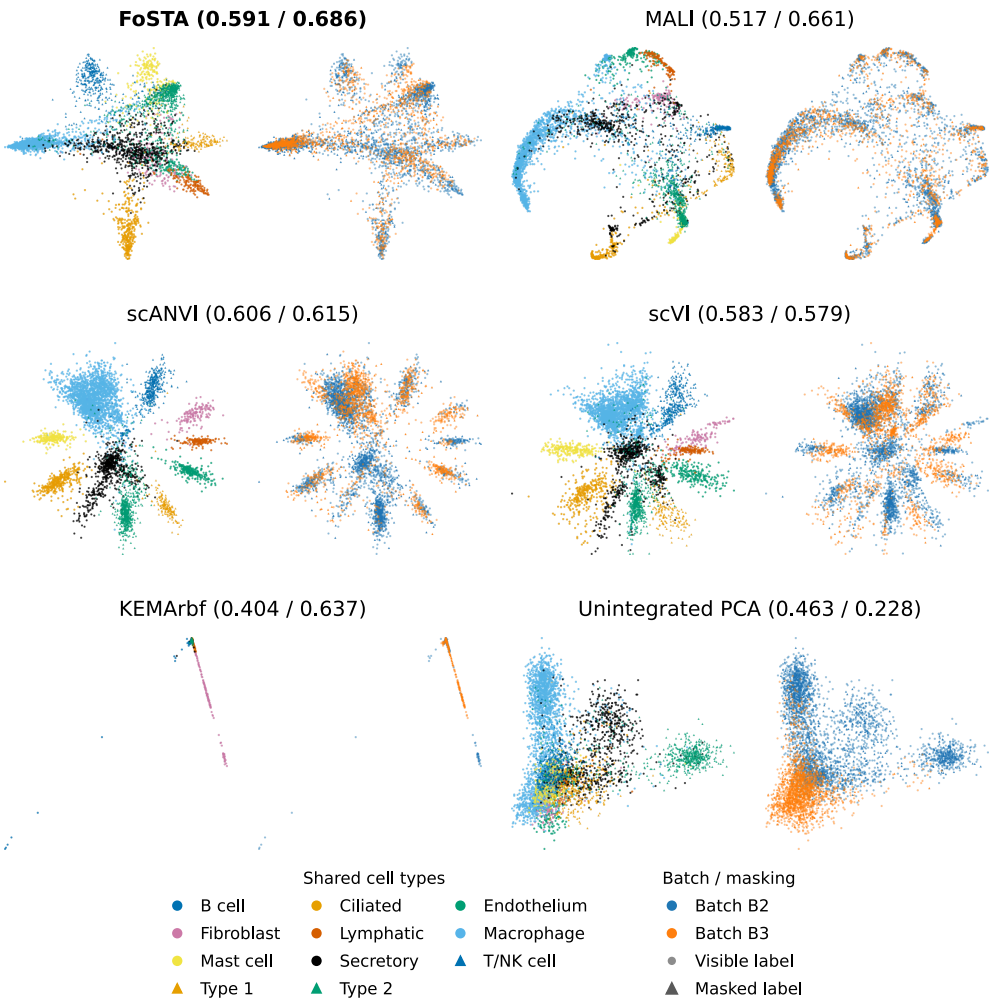

Figure 4: **Pairwise integration visualizations for the B2/B3 real batch pair from the lung dataset [51] using seed 56089.** For each method, the left panel is labeled by cell type and the right panel by batch; values in parentheses report biological conservation (Bio) and batch correction (Batch) scores, respectively. The selected example is representative of the aggregated quantitative results across the B-family batches and seeds. FoSTA produces a structured embedding with meaningful local groupings and good domain mixing while clearly distinguishing cell types. MALI also achieves good batch mixing and preserves global structure, but cell types are less clearly separated. scANVI and scVI produce more isolated cluster configurations, while KEMArbf strongly compresses the embedding and loses biological structure. The unintegrated PCA embedding illustrates the original batch effects present in the data.

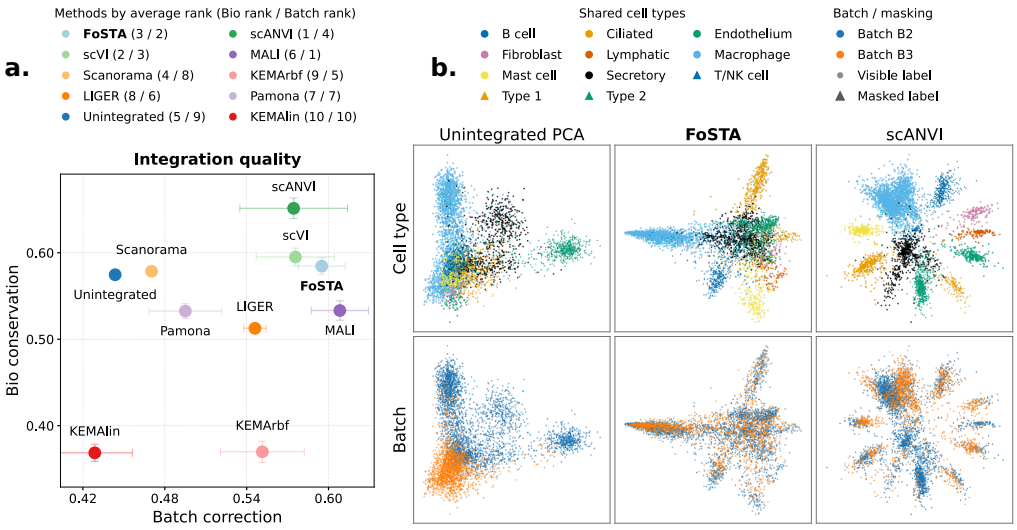

Figure 5: **Results of batch effect removal on lung single-cell data [50] under reduced label availability, with 50% of labels masked instead of 20% as in Fig. 2. a.** Biological preservation (Bio) vs. batch correction (Batch) scores [51], averaged across 36 batch pairs and five seeds. Error bars represent seed variability, computed as the average within-pair standard deviation across the 36 batch pairs. **b.** Embeddings of the simulated batch integration task. Left column: data before correcting for batch effects, visualized with PCA. Center column: resulting embedding after applying FoSTA to correct the batch effects. Right column: resulting embedding after applying scANVI to correct the batch effects. The plots are colored by cell type (top) and batch (bottom). FoSTA mixes the batches well while preserving the biological structure of the data.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction accurately reflect the main contributions and scope of the paper. These claims are supported in the main text and appendix through methodological descriptions and empirical results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly discuss the limitations of our method in the discussion section (Section 5).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions and proofs of Lemmas 1 and 2 are provided in Appendices A and C, respectively.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides sufficient details to reproduce the main experimental results, including a description of the model architecture (Section 3 and Fig. 1), implementation and hyperparameter settings (Appendix F), and dataset, preprocessing, experimental protocol, and evaluation details for the real-world multimodal, synthetic, and single-cell experiments (Appendices G, H, and I).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code to reproduce all experiments is publicly available at `https://github.com/aumona/FoSTA` and is linked directly in Section 4 of the main paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the relevant experimental details, including datasets, preprocessing, experimental protocols, implementation and hyperparameter settings, and evaluation metrics. These details are provided in the main text (Section 4) and expanded for the real-world multimodal, synthetic, and single-cell experiments in Appendices F, G, H, and I.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports standard deviations for the main experimental results across multiple random seeds, including the real-world multimodal experiments in Section 4.1 and the single-cell batch alignment experiments in Section 4.2, as well as the extended experiments reported in Appendices G, H, and I.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the compute resources required to reproduce the experiments. These details are documented in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research presented in the paper fully complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses both the potential positive and negative societal impacts of the proposed work in the Broader Impacts section.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not involve models or datasets that pose a high risk of misuse or dual use. The proposed method is intended for scientific research and does not generate or process sensitive or potentially harmful content.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All external datasets and models used in the paper are properly cited with references to their original sources in Section 4 and Appendices G.1, H.1 and I.1.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The FoSTA source code and experimental pipelines are publicly available, together with documentation, installation instructions, and scripts to reproduce the experiments. The repository is linked in Section 4.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper uses a publicly available single-cell dataset that was collected and shared by third parties. We do not conduct any new research involving human subjects or crowdsourcing, and the data used is ethically sourced.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve any direct research with human subjects or participant interaction. All data used are publicly available and were collected under proper ethical oversight by the original data providers.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: No LLMs were used in the development of the core methods presented in this research. See the AI Use Disclosure section for details.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.