MIND YOUR MARGINALS: WHAT GOVERNS FIDELITY IN LINEARIZED HIERARCHICAL OPTIMAL TRANSPORT?

Anonymous authors
Paper under double-blind review

ABSTRACT

Linearized optimal transport (LOT) replaces the $\binom{n}{2}$ transport solves needed to compare n measures with n offline embeddings into a tangent space at a fixed reference. We study LOT for *hierarchical* transport, where the ground metric between support atoms is itself a Wasserstein distance over a learned, coarsened support, and instantiate this construction (HILOT) in text, graphs and images from one shared implementation. Our central finding is that fidelity is governed less by modality than by four construction-level choices, the first of which has been overlooked. **(i) Orientation:** a transport plan between measures on a shared atom set admits two barycentric projections, indexed by the measure or by the reference, and only the latter lies in a tangent space shared across objects. Across 24 arms in two modalities the benefit of correcting the orientation is proportional to the linearization gap it repairs (Spearman $\rho = -0.946$, $p = 3 \times 10^{-12}$; $\rho = -0.947$ and -0.806 within graphs and images), reducing test error by up to 0.235 on a single benchmark. **(ii) Plan density:** an exact plan is a polytope vertex with at most $n + m - 1$ nonzeros, so the embedding degenerates into a quantizer; entropic regularization removes this with a provably interior optimum. Orientation and regularization together close the gap to exact hierarchical transport on curated retrieval (0.1123 against 0.1121 NDCG@10, $t = -0.04$) at ~ 100 queries/s against 1.5. **(iii) Coarsening** acts as capacity control, with a truncation sweep isolating coarsening rather than support truncation as the mechanism behind its one clear failure case. **(iv) The component model** is exchangeable and matters about as much as coarsening, an order of magnitude less than orientation. Across six text benchmarks the linearization preserves accuracy relative to exact hierarchical transport at 12 to 182 times its throughput.

1 INTRODUCTION

Documents as distributions over topics, graphs as distributions over structural roles, images as distributions over patch descriptors: optimal transport (OT) supplies a metric for such objects that respects support geometry rather than treating coordinates as exchangeable features (Villani, 2009; Peyré & Cuturi, 2019). Cost is the obstacle. An exact Wasserstein distance requires solving a linear program, and comparing n objects requires $\binom{n}{2}$ solves. Linearized optimal transport (Wang et al., 2013) fixes a reference σ, computes one transport map from σ to each object, and uses the barycentric projection of that map as a Euclidean embedding, reducing $\binom{n}{2}$ solves to n offline embeddings followed by nearest-neighbour search. Distortion bounds (Mérigot et al., 2020; Moosmüller & Cloninger, 2023; Delalande & Mérigot, 2023) control the error in terms of support and reference geometry.

We study LOT in a setting the original formulation does not cover: *hierarchical* transport, where the ground metric between atoms is itself a Wasserstein distance over a finer space. In our construction, HILOT, level-0 atoms occupy $\mathbb{R}^d$ with a Euclidean ground cost, and level-1 *components* are sparse distributions over atoms whose pairwise cost is the W_1 distance between those distributions. Objects are mixtures over components. Linearizing at level 1 is not a reparameterization of flat LOT: the ground cost is itself a transport distance, and the support is learned and coarsened rather than given.

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

Instantiating for text recovers hierarchical optimal topic transport (Yurochkin et al., 2019) at level 1 and linearizes it (§6.1), but that is the entry point rather than the contribution. Our substantive finding is that linearization *fidelity* is governed by four choices that cut across modality, and that one of them, which marginal indexes the barycentric projection, dominates every modelling decision usually reported. One implementation instantiated in three modalities separates modality-specific effects from construction-level ones.

Contributions

1. A domain-agnostic hierarchical LOT construction with a composition result showing that at $p = 1$ the two-level object is a metric measure space whose ground metric is itself a Wasserstein distance (Prop. 1), instantiated for text, graphs and images from one core.
2. An orientation law (§4.1, §6.2): we formalize the two admissible projections, show only the reference-indexed one lies in a shared Hilbert space, and establish across 24 arms that the benefit of correcting orientation scales with the gap it repairs.
3. A degeneracy analysis of the exact plan with matching entropic limits (Props. 2, 3), predicting and confirming an interior optimum in regularization strength.
4. A scope characterization of coarsening and of the component model, including a truncation sweep separating coarsening from truncation.

We do not claim linearized graph OT as novel (Kolouri et al., 2021; Beier et al., 2022); our flat arm reimplements that shape as the control against which the hierarchical arm is measured.

2 Related Work

Word Mover's Distance (Kusner et al., 2015) compares bag-of-words measures over word embeddings under W_1, motivating relaxations and random-feature approximations (Wu et al., 2018). Hierarchical Optimal Topic Transport (Yurochkin et al., 2019) coarsens the support to LDA topics (Blei et al., 2003; Griffiths & Steyvers, 2004) and uses topic-to-topic transport as ground cost, but remains pairwise. On graphs, Wasserstein Weisfeiler–Leman (Togninalli et al., 2019) compares distributions over WL-refined embeddings (Shervashidze et al., 2011) under a Euclidean (flat) ground cost; Gromov–Wasserstein formulations (Mémoli, 2011; Vayer et al., 2019) drop the shared-support assumption. LOT (Wang et al., 2013) was introduced for continuous image measures, with distortion bounds established later (Mérigot et al., 2020; Moosmüller & Cloninger, 2023; Delalande & Mérigot, 2023) and supervised variants in Khurana et al. (2023); WEGL (Kolouri et al., 2021) linearizes graph comparison and Beier et al. (2022) give a linearized Gromov–Wasserstein distance. To our knowledge, which marginal indexes the barycentric projection has not been isolated as a design variable, nor has hierarchical linearization been studied across modalities.

3 The HiLOT Construction

For $\mu \in \Delta_n$, $\nu \in \Delta_m$ and cost $C \in \mathbb{R}^{n\times m}$, write $\Gamma(\mu,\nu) = \{\Pi \geq 0 : \Pi \mathbf{1}_m = \mu,\ \Pi^\top \mathbf{1}_n = \nu\}$ and $W_1(\mu,\nu;C) = \min_{\Pi\in\Gamma(\mu,\nu)} \langle C, \Pi\rangle$. When C is a metric on a common ground set this is the Kantorovich–Rubinstein distance and is itself a metric on the simplex.

Two-level geometry Let atoms $\mathcal{A} = \{a_1, \dots, a_K\}$ have coordinates $A \in \mathbb{R}^{K\times d}$ and level-0 cost $C^{(0)}_{jk} = \|A_j - A_k\|_2^p$. Let components $\{\bar{c}_s\}_{s=1}^S \subset \Delta_K$ be obtained by retaining the m highest-mass atoms of each and renormalizing. Define

$$e_s = A^\top \bar{c}_s \in \mathbb{R}^d, \qquad C^{(1)}_{st} = W_1\big(\bar{c}_s, \bar{c}_t;\, C^{(0)}\big), \tag{1}$$

the component coordinate and the **nested** ground cost. An object is a mixture $\mu \in \Delta_S$, and the exact hierarchical distance is $W_1(\mu,\nu;C^{(1)})$. The *flat* variant uses atoms directly with $C^{(0)}$: this is the prior-art shape (Kolouri et al., 2021; Togninalli et al., 2019) and the control throughout. Write $E \in \mathbb{R}^{S\times d}$ for the component coordinate matrix.

Proposition 1 (Metric composition)**.** *Let $p = 1$. Then $W_1(\cdot,\cdot;C^{(0)})$ is a metric on Δ_K; $C^{(1)}$ is a pseudometric on $\{1,\dots,S\}$, and a metric iff $s \mapsto \bar{c}_s$ is injective; hence $W_1(\cdot,\cdot;C^{(1)})$ is a*

pseudometric on Δ_S, and a metric whenever $C^{(1)}$ is. For $p > 1$ the first statement fails, since $\|\cdot\|_2^p$ violates the triangle inequality.

Proofs are in Appendix A. HILOT therefore defines a metric measure space whose ground metric is itself a Wasserstein distance, which distinguishes it from a flat linearization on component barycenters. Note also that HOTT-style pairwise transport truncates component proportions below $1/(S+1)$ before solving, which is not measure preserving and voids the third statement; our linearized arms need no truncation, so the linearization is in this respect more geometrically faithful than the baseline it replaces.

Linearization and the reference Fix $\sigma \in \Delta_S$ on the component index space, a Gaussian mixture with k evenly spaced centres and full-width-at-half-maximum (FWHM) controlling overlap: the i-th centre sits at index $\lfloor (i-1)(S-1)/(k-1) \rfloor$, each component is a Gaussian in index units, and the mixture is normalized (Fig. 6, left, Appendix D). For an object μ let

$$\Pi^\mu \in \operatorname*{arg\,min}_{\Pi \in \Gamma(\mu,\sigma)} \langle C^{(1)}, \Pi \rangle, \tag{2}$$

so $\Pi^\mu \mathbf{1} = \mu$ (rows) and $(\Pi^\mu)^\top \mathbf{1} = \sigma$ (columns). Total cost is n solves offline plus Euclidean comparisons against $\binom{n}{2}$ solves, and the vectors are compatible with approximate nearest-neighbour indices (Johnson et al., 2019; Malkov & Yashunin, 2020).

Instantiations **Text**: atoms are vocabulary words at GloVe coordinates (Pennington et al., 2014); components are topics represented by their 20 highest-weighted words, renormalized, so $C^{(1)}$ is transport between topic word-distributions. We write LOTT-k for k reference components. **Graphs**: atoms are Weisfeiler–Leman labels after r refinement rounds, coordinatized by mean attribute where available and otherwise a standardized ten-dimensional structural descriptor, truncated to the K most frequent labels. Components (*motifs*) are coordinate clusters; WWL/H-WWL denote exact pairwise flat/hierarchical transport, L-WWL/L-H-WWL their linearizations. **Images**: atoms are visual words from a codebook over gradient-orientation patch descriptors (low-energy patches suppressed); components are visual topics from the same clustering routine as graph motifs. Full specifications in Appendix B.

4 Four Determinants of Fidelity

4.1 Orientation: which marginal indexes the tangent space

Both marginals of Π^μ live on the *same* S components with the same coordinates E, so the barycentric projection admits two realizations:

$$\Phi_\sigma^{\text{SRC}}(\mu)_i = \frac{1}{\mu_i} \sum_s \Pi^\mu_{is}\, e_s, \qquad \Phi_\sigma^{\text{REF}}(\mu)_i = \frac{1}{\sigma_i} \sum_s \Pi^\mu_{si}\, e_s, \tag{3}$$

with the embedding the concatenation over $i = 1, \ldots, S$ in either case. These are not interchangeable. The second is the projection of the plan read as a map $\sigma \to \mu$, the discrete realization of $F_\sigma(\mu) = T_\sigma^\mu$ as an element of $L^2(\mathbb{R}^S, \sigma)$ (Wang et al., 2013; Mérigot et al., 2020): every object is weighted by the same σ and indexed by the same basepoint, so all embeddings lie in one Hilbert space and Euclidean distance is a comparison within a single tangent space. The first is the projection of the reverse map and lies in $L^2(\mathbb{R}^S, \mu)$, which *varies with the object*, with two practical consequences: the weighting differs between any two embeddings, so their Euclidean distance is not an inner-product comparison in a common space; and whenever $\operatorname{supp}\mu \subsetneq \{1, \ldots, S\}$ the blocks indexed by $i \notin \operatorname{supp}\mu$ are undefined and must be assigned by convention, which we take to be zero. The support pattern therefore depends on μ and most blocks vanish for sparse μ, whereas σ has full support by construction. We treat orientation as an explicit design variable and also evaluate Φ^{BOTH}, the concatenation of both with each half normalized to unit norm; since both projections are read off the *same* plan this costs no additional transport, only a doubled search dimension.

4.2 Plan density: vertex sparsity and its entropic remedy

Proposition 2 (Quantization of the exact plan)**.** *Let Π be an optimal basic feasible solution of* (2) *with $n' = |\operatorname{supp}\mu|$, $m' = |\operatorname{supp}\sigma|$. Then $\|\Pi\|_0 \le n' + m' - 1$, so at most $m' - 1$ rows contain*

more than one nonzero and for every singleton row i the block $\Phi_\sigma^{\mathrm{SRC}}(\mu)_i$ reduces to $e_{s(i)}$: a hard assignment rather than an average.

The bound is attained exactly. With $S = 70$ atoms and full-support marginals we measure a median of $139 = 2 \cdot 70 - 1$ nonzeros, 1.99 nonzeros per row, and 85 to 91% singleton rows for both documents and queries. The linearization therefore behaves as a near-permutation matching rather than a smooth projection, and this is a property of the polytope rather than of how concentrated μ is. The entropic relaxation (Cuturi, 2013)

$$\Pi^{\mu,\varepsilon} = \operatorname*{arg\,min}_{\Pi \in \Gamma(\mu,\sigma)} \langle C^{(1)}, \Pi \rangle - \varepsilon H(\Pi), \quad H(\Pi) = -\sum_{is} \Pi_{is} \log \Pi_{is}, \tag{4}$$

gives a strictly positive plan for every $\varepsilon > 0$. Both limits are degenerate, and asymmetrically so.

Proposition 3 (Limits of the entropic embedding)**.** *As $\varepsilon \to 0^+$, $\Pi^{\mu,\varepsilon}$ converges to a solution of* (2) *and Prop. 2 applies. As $\varepsilon \to \infty$, $\Pi^{\mu,\varepsilon} \to \mu\sigma^\top$ and $\Phi_\sigma^{\mathrm{SRC}}(\mu)_i \to \sum_s \sigma_s e_s$ while $\Phi_\sigma^{\mathrm{REF}}(\mu)_i \to \sum_s \mu_s e_s$ for every i. Both limits are constant in i, but the source-indexed limit does not involve μ and therefore carries no information, whereas the reference-indexed limit is the barycentric projection of μ itself.*

This predicts an interior optimum in ε, and predicts over-regularization to be far more damaging under source-indexing; both are confirmed in §6.3. Numerically, (4) must be solved in the log domain (Schmitzer, 2019) with ε expressed as a fraction of the cost scale, since (1) has entries of order 1 to 10.

4.3 Coarsening, the component model, and the bounds

The nested cost is defined over $S \ll K$ components, so HiLOT discards resolution while gaining a ground metric with distributional structure. This is a bias-variance trade, predicting that the hierarchical arm helps when atom coordinates are weak or data scarce and hurts when atoms are individually discriminative and data abundant. Separately, the *component model* that produces $\{\bar{c}_s\}$ is exchangeable without touching the nested cost, the reference, or the linearization, giving a fourth axis; we compare clustering over atom coordinates, a generative mixture over atom counts (Blei et al., 2003), and object clustering with class-based weighting after Grootendorst (2022) and Zhang et al. (2022), at matched component count.

The available guarantees, e.g. $W_2(\mu,\nu) \le \|T_\mu - T_\nu\|_{L^2(\rho)} \le C\, W_p(\mu,\nu)^{2/15}$ for ρ Lebesgue on compact convex $X \subset \mathbb{R}^d$ (Mérigot et al., 2020), assume absolutely continuous measures and do not apply verbatim to discrete measures on a finite metric space. We invoke them as guidance: the bound is one-sided, its exponent weak, its constant ground-geometry dependent, and it is stated for the reference-indexed embedding — each feature has an empirical counterpart below.

5 Protocol

Classification uses k-NN with $k = 7$. On graphs and images we use repeated random $75/25$ splits with 10 repeats, reporting mean$\pm$std and, where two arms see identical splits, *paired* differences, with $|t| < 2$ indicating no detectable difference. Pairwise matrices are split-independent, computed once and sliced per split. Timing is attributed by share of *unique* OT solves ($n(n-1)/2$ per symmetric matrix), since attributing by entries would double-credit the pairwise baseline. Retrieval reports query-time throughput, queries divided by query embedding plus search time, with corpus embedding and index construction reported separately. Text uses six classification benchmarks (Table 6, Appendix B) and BEIR (Thakur et al., 2021) for retrieval with curated judgements; graphs are TU datasets (Morris et al., 2020), $n = 188$ to 4337; images are MNIST, Fashion-MNIST and USPS at 200 per class.

Our graph and image protocols are not comparable to published kernel or convolutional numbers, so the controlled comparison throughout is linearized against our own pairwise under an identical protocol. Text classification timings come from a consumer CPU while all other timings come from cluster CPU nodes (Appendix B), so throughput is normalized within platform and absolute wall-clock figures are never placed on a common axis.

Table 1: Text instantiation. Left: mean test error. Right: throughput normalized so exact pairwise HOTT= 1.0. All reference counts and the HOFTT ablation appear in Appendix C.

	Test error				Relative throughput		
Dataset	HOTT	LOTT-1	LOTT-5	LOTT-10	LOTT-1	LOTT-5	LOTT-10
BBCSPORT	**0.036**	0.164	0.079	0.070	11.5	**12.7**	12.5
TWITTER	0.350	0.342	0.331	**0.321**	98.6	**101.8**	101.1
OHSUMED	**0.462**	0.523	0.478	**0.462**	**136.0**	125.8	124.4
CLASSIC	**0.051**	0.082	0.057	0.053	**93.3**	87.7	86.2
REUTERS	0.062	0.113	0.103	0.093	**92.0**	91.4	90.0
AMAZON	**0.110**	0.133	0.114	0.112	**182.3**	168.0	168.4

6 RESULTS

6.1 TEXT INSTANTIATION

Table 1 reports test error and throughput on the six text benchmarks. Linearization preserves accuracy relative to exact pairwise HOTT at 12 to 182 times its throughput, and on TWITTER the linearized variant is the strongest method in the family. Accuracy improves monotonically with reference component count on all six corpora and the method remains stable under reduced supervision (Appendix D). Against transformer encoders on the same CPU, the strongest sentence encoder attains lower error at roughly 196 times the inference cost, while a lighter sentence encoder gives 0.051 lower average error at 33 times the cost; against DistilBERT, RoBERTa and BERT the linearized embedding is competitive on error and dominant on cost (Appendix C). Interpolating LOTT-10 distances with BM25 scores (Robertson & Zaragoza, 2009) reduces error on five of six corpora, sharply on BBCSPORT (0.070 to 0.027) and REUTERS (0.093 to 0.040), indicating that topic geometry captures structure lexical overlap does not.

Under t-SNE (Fig. 6, Appendix D) the linearized embedding separates the four CLASSIC classes into distinct, internally homogeneous clusters where a normalized bag-of-words baseline barely does; exact transport gives tighter inter-class separation but less intra-class cohesion.

6.2 ORIENTATION DOMINATES, AND ITS BENEFIT SCALES WITH THE GAP

Figure 1 plots, for each dataset-level arm, the *linearization gap* of the default source-indexed embedding, its test error minus that of exact pairwise transport on the same splits, against the paired improvement from switching to reference-indexing. Across 24 arms the relationship is monotone and tight (Spearman $\rho = -0.946$, $p = 3 \times 10^{-12}$) and replicates within each modality separately ($\rho = -0.947$, $p = 2.8 \times 10^{-7}$ on 14 graph arms; $\rho = -0.806$, $p = 0.0049$ on 10 image arms), so it is not an artifact of pooling. Of the 19 arms with $|\text{gap}| > 0.01$, 18 have the predicted sign. Figure 2 shows the mechanism: source-indexing activates between 4.5 and 78.5 of 174 to 300 barycentric blocks, reference-indexing all of them.

Effect sizes are large (Table 2). On MNIST/flat, reference-indexing reduces error from 0.4258 to 0.1910 ($t = -39.9$), recovering 85% of a gap of $+0.278$; on MUTAG/flat, from 0.2489 to 0.1468 ($t = -6.00$), recovering 70% of $+0.147$. Where the source-indexed embedding already matched exact transport, on ENZYMES and PROTEINS_full with attribute coordinates and on NCI1, reference-indexing degrades accuracy by $+0.016$ to $+0.078$. For calibration, the largest orientation effect of 0.235 is roughly five times the largest flat-versus-hierarchical effect we measure. Φ^{BOTH} is the best fixed orientation on 11 of the 14 arms where all three are available, at no additional transport cost. Because choosing per dataset on test error would constitute test-set tuning, we also select per split on a 20% train-internal validation split, which beats the default on 11 of 14 arms at a cost of 0.007 to 0.017 against the oracle. Two negative observations: PROTEINS/flat violates the law (gap $= +0.0115$, $\Delta = +0.0287$, $t = +2.84$), sitting at the sign-flip threshold with one-dimensional attribute coordinates and hence a nearly degenerate ground cost; and on IMDB-MULTI the reference-indexed embedding has a single distinct block of 300, so that dataset is reported for completeness only. Fixing the orientation also resolves a cross-modal discrepancy: increasing the reference count improves accuracy on all six text corpora but degrades it on images under source-indexing (0.3118 to 0.4258 for $k = 1, 5, 10$ on MNIST/flat), whereas under reference-indexing the trend is restored (0.2426 to 0.1910).

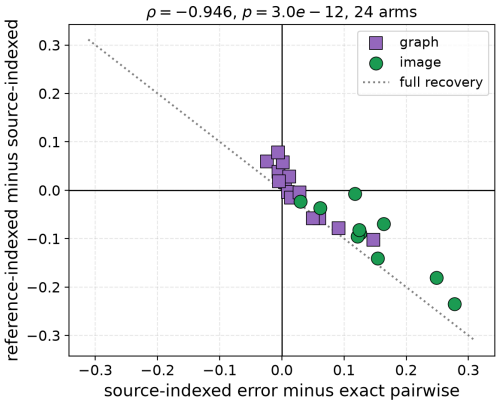


Figure 1: The orientation law across 24 arms in two modalities. Reference-indexing is a corrective whose benefit scales with the loss it repairs, and it adds error where no gap exists.

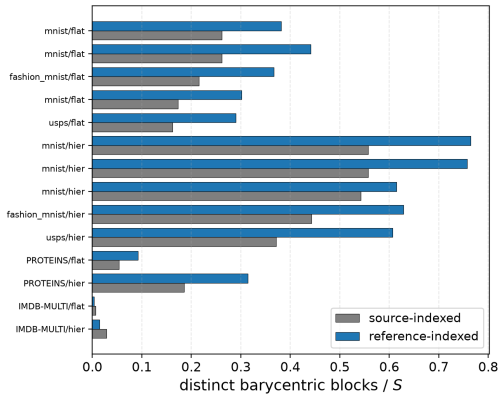


Figure 2: Distinct barycentric blocks as a fraction of S. The source-indexed projection discards most of the embedding whenever the measure is sparse.

Table 2: Orientation on representative arms, ordered by linearization gap. Bold marks the best linearized arm. Φ^{BOTH} concatenates both projections of the same plan at no extra transport cost; *sel.* selects per split on a train-internal validation set. All 24 arms appear in Table 19.

Arm	mod.	exact	Φ^{SRC}	Φ^{REF}	Φ^{BOTH}	sel.	gap	$\Delta(\text{REF}-\text{SRC})$
MNIST / flat	image	0.148	0.426	0.191	**0.184**	0.189	$+0.278$	-0.235 (t=-39.9)
USPS / flat	image	0.096	0.250	0.109	**0.105**	0.106	$+0.154$	-0.141 (t=-51.5)
MUTAG / flat	graph	0.102	0.249	**0.147**	—	—	$+0.147$	-0.102 (t=-6.0)
MUTAG / hier	graph	0.109	0.200	**0.121**	—	—	$+0.092$	-0.079 (t=-5.5)
F-MNIST / flat	image	0.264	0.325	0.289	**0.261**	0.265	$+0.061$	-0.037 (t=-11.0)
DD / hier	graph	0.391	0.440	**0.382**	—	—	$+0.050$	-0.058 (t=-6.2)
PROTEINS / flat	graph	0.296	0.307	0.336	**0.298**	0.309	$+0.012$	$+0.029$ (t=$+2.8$)
ENZYMES / hier	graph	0.689	**0.683**	0.761	—	—	-0.006	$+0.078$ (t=$+7.0$)
NCI1 / flat	graph	0.244	**0.219**	0.279	—	—	-0.024	$+0.059$ (t=$+10.5$)

6.3 Plan density: an interior optimum, as predicted

Sweeping entropic strength on curated retrieval (Fig. 3) confirms Props. 2 and 3: quality rises from the exact-plan value as the plan densifies and falls as it approaches the independent coupling, peaking at $\varepsilon \approx 0.15$ to 0.20 of the cost scale. At the optimum the entropic plan recovers roughly half of a gap that is itself significant (exact HOTT minus source-indexed LOTT-10: $+0.0260 \pm 0.0092$ NDCG@10, $t = +2.82$, $p = 0.005$ on NFCORPUS), and mean nonzeros per row rise from 1.99 to S. Marginal violations stay below 1.5×10^{-10} throughout, so the sweep measures regularization rather than solver failure; with seven settings compared we treat individual p-values as descriptive and rely on replication.

Combining the two determinants closes the gap entirely (Table 3). On NFCORPUS, reference-indexing at $\varepsilon = 0.20$ attains NDCG@10 $= 0.1123$ against 0.1121 for exact pairwise HOTT ($t = -0.04$) at ~ 100 queries/s against 1.5; on SCIFACT, 0.1514 against 0.1533 ($t = +0.18$). The decomposition differs between corpora, orientation contributing most on NFCORPUS and regularization most on SCIFACT, so we report the combination rather than either ingredient alone. Across the transport family as a whole, exact included, BM25 and a sentence encoder outperform every arm on BEIR by a wide margin; the claim here concerns the cost of computing hierarchical transport, not the standing of topic geometry against modern retrievers.

Remark 1 (A hypothesis tested and rejected)**.** Query concentration is a natural explanation for the retrieval gap (query posteriors have median effective support 3.7 of 70 atoms against 9.9 for documents), but the per-query correlation between effective support and the quality gap is null (Spearman $\rho = -0.016$, $p = 0.80$) and the gap is flat across concentration terciles. Prop. 2 explains why: the vertex bound is independent of $|\operatorname{supp}\mu|$. A metric-diffusion remedy raising effective support recovers only $+0.0034$ NDCG@10, within noise, whereas entropic regularization on the plan recovers $+0.0134$ (Appendix F).

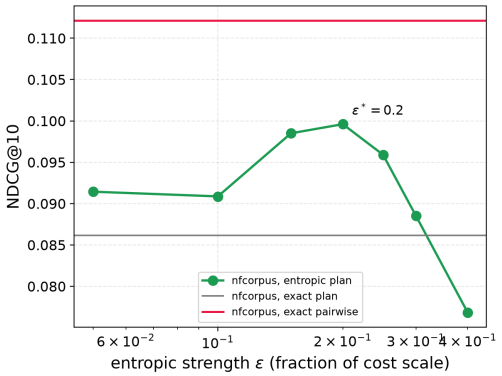


Figure 3: Entropic de-quantization on curated retrieval. Quality peaks at an interior ε, as Prop. 3 predicts.

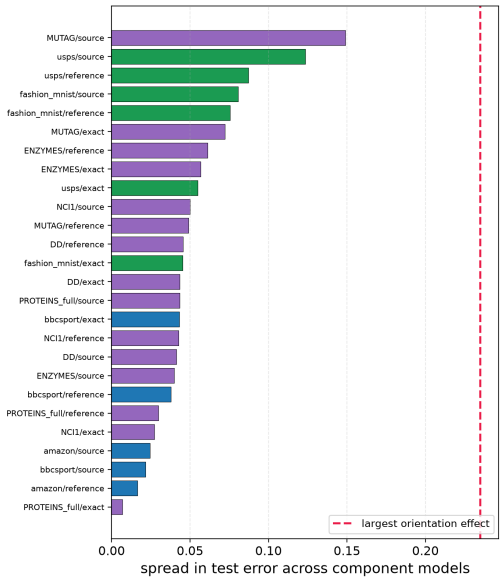


Figure 4: Spread in test error across component models per arm, against the largest orientation effect. The component model is a secondary axis.

Table 3: Curated retrieval, NDCG@10 and query-time throughput. Orientation and entropic regularization together close the gap to exact hierarchical transport on both corpora. BM25 and a sentence encoder are context rows.

	NFCORPUS ($\lvert C\rvert$=3633)		SCIFACT ($\lvert C\rvert$=5183)	
Method	NDCG@10	queries/s	NDCG@10	queries/s
HOTT (exact pairwise)	0.1121	1.5	0.1533	1.1
LOTT-1, Φ^{SRC}	0.0418	549	0.0558	559
LOTT-5, Φ^{SRC}	0.0792	493	0.1006	500
LOTT-10, Φ^{SRC}	0.0855	503	0.1239	488
LOTT-10, Φ^{SRC}, ε=0.2	0.0996	94	—	—
LOTT-10, Φ^{REF}	0.1075	480	0.1282	470
LOTT-10, Φ^{REF}, ε=0.2	**0.1123**	94	**0.1514**	92
BM25 (context)	0.3231	8605	0.5454	761
Sentence encoder (context)	0.3388	1671	0.6393	831

6.4 Coarsening is capacity control, and the component model is secondary

At matched reference size the nested ground cost significantly helps on MUTAG ($\Delta = -0.049$, $t = -3.2$), DD (-0.043, $t = -3.7$ at $k = 1$) and PROTEINS (-0.036, $t = -3.1$ at $k = 1$), is null on ENZYMES, PROTEINS_full and IMDB-MULTI, and significantly hurts on NCI1 ($+0.042$, $t = +8.2$), while running 2 to 3.5 times faster because motif mixtures live on $S = 70$ atoms against $K = 300$ to 500 WL labels. A truncation sweep identifies the cause of the NCI1 exception: holding $S = 70$ and raising K from 300 to 3000, the hierarchical penalty grows monotonically from $+0.042$ to $+0.103$ ($t = +8.2$ to $+20.3$) while the flat arm is flat in K. The loss is therefore attributable to *coarsening* rather than to truncation of the atom set. We note the WL-subtree kernel also improves with K and overtakes the flat linearization for $K \geq 1000$, so that ordering holds only at $K = 300$.

Two ablations bear on ground-metric quality. Replacing informative structural coordinates with Gaussian random coordinates on MUTAG degrades every arm including exact pairwise transport (0.102 to 0.177), the empirical counterpart of the geometry-dependent constant in the bounds. A four-feature descriptor also collapsed 174 MUTAG WL labels onto four distinct coordinate rows, capping clustering at four components and making a component-count ablation appear spuriously

Table 4: Component models at matched count $S = 70$, reference-indexed, $k = 10$. Cluster is over atom coordinates; generative is a mixture over atom counts; object clustering uses class-based weighting. Setup is the one-time component fit. †Soft codeword assignment disadvantages the generative model on images because it requires integral counts; see Table 21.

		Test error			Setup (s)		
Dataset	mod.	cluster	generative	object	cluster	generative	object
MUTAG	graph	**0.121**	0.170	**0.121**	1.5	4.4	1.3
ENZYMES	graph	0.761	0.735	**0.700**	1.1	5.6	1.4
PROTEINS_full	graph	0.339	0.346	**0.317**	1.6	25.9	2.4
DD	graph	**0.382**	0.428	0.412	24.8	29.3	25.0
NCI1	graph	0.277	0.285	**0.242**	11.5	68.4	12.2
USPS	image	0.145	0.349†	**0.120**	7.8	18.4	8.2
F-MNIST	image	**0.316**	0.391†	0.332	8.5	34.9	8.4
BBCSPORT	text	—	**0.032**	0.070	—	52.5	52.0
AMAZON	text	—	0.093	**0.077**	—	619.7	478.5

insensitive; a ten-feature descriptor gives 102 distinct rows and 70 effective components, so we report effective component count alongside requested count throughout.

Comparing three component-model families at matched $S = 70$ on identical splits (Table 4) yields two findings. The choice is consequential but secondary: the median spread in test error is 0.044 over 26 arms, comparable to the flat-versus-hierarchical effect and roughly a fifth of the largest orientation effect (Fig. 4). And contrary to the assumption that a generative mixture is the natural default, it is the weakest choice on a majority of arms while being three to six times more expensive to fit. One caveat is load bearing: the generative model requires integral counts, and soft codeword assignment on images distributes each patch's unit mass as fractions, so rounding discards roughly half the retained mass; a hard-assignment control on USPS restores the retained mass from 8.5 to 18.0 counts per image (Table 21). The component model also interacts with orientation: object clustering on MUTAG is catastrophic under source-indexing ($+0.149$, $t = +9.04$) yet ties for best under reference-indexing; and on USPS the best component model under reference-indexing (0.120) beats the generative model computed with *exact* transport (0.299), so the component model can matter more than whether transport is computed exactly.

6.5 Retrieval and scaling

Retrieval is quadratic for pairwise OT by construction, since every query is scored against every corpus object, whereas the linearized arm amortizes transport offline. On graphs, paired over shared query splits, the flat linearization is significantly *better* than exact pairwise transport on NCI1 (MAP $+0.0111$, $t = +4.14$) and ENZYMES (MAP $+0.0329$, $t = +3.94$), better on P@1 on Mutagenicity ($+0.0124$, $t = +3.19$), and never significantly worse, at 124–358 queries/s against 0.8–3.1. On Fashion-MNIST the concatenated orientation beats exact hierarchical transport (-0.0132, $t = -2.41$), a margin validation-based selection retains (-0.0100, $t = -2.05$). Relevance is same-class membership, so we report P@k, MRR, MAP and R-precision rather than recall at small cutoffs.

Pairwise throughput is n-independent (constant cost per solve); linearized cost is an for embedding plus bn^2 for search, so throughput $\propto n/(a{+}bn)$, growing with n and saturating at $1/b$. Both halves hold: exact throughput is $3232{\pm}100$ solves/s across a 22-fold range of n (188–4110), while linearized speedup rises from $51.8\times$ to $1115\times$ over the same range (Appendix C).

6.6 Basepoint selection

The reference is the linearization basepoint and its placement is a geometric choice. Increasing component count improves accuracy monotonically on all six text corpora and, once orientation is corrected, on images as well. Widening FWHM from 10 to 1000 reduces test error from 0.540 to 0.447 on OHSUMED at essentially unchanged throughput, while corpora with at most five classes vary by under 0.006; a practical default is FWHM $\approx S/k$. The useful range is narrower on graphs, where on ENZYMES the sweep alone spans 0.653 to 0.760 test error, wider than the entire

hierarchical-versus-flat effect. Since the default reference has effective support 69.9 of 70, FWHM shapes the plan less than its name suggests (Appendix C).

7 DISCUSSION

Modality matters less than four construction-level choices. Orientation is largest: its effect follows a quantitative law across 24 arms and two modalities, and the empirically superior convention is also the formally correct one. Plan density is second: an exact plan is a polytope vertex, so the linearization is intrinsically a quantizer, and entropic regularization fixes this with a provably interior optimum. Coarsening is third and behaves as capacity control. The component model is fourth and, contrary to the usual assumption, a generative mixture is typically the weakest and most expensive option. With the determinants set correctly, the linearization matches or exceeds exact transport on retrieval at three to four orders of magnitude greater query-time throughput, including four arms where it is significantly better, tentatively attributable to the smoothing effect of barycentric projection on noisy per-pair costs.

Limitations Graph and image protocols are not comparable to published kernel or convolutional baselines, and text classification timings come from a different CPU. Corpora reach 9152 documents and 4337 graphs, so the $O(n^2)$-to-$O(n)$ claim spans roughly 1.5 decades of n. Graph and image relevance is same-class membership; only text retrieval uses curated judgements. The orientation law rests on 24 arms with one genuine violation and is partly mechanical (an arm with no gap has no headroom), so it characterizes rather than predicts. IMDB-MULTI is degenerate under both orientations, and two text corpora were excluded from the component-model comparison for unverifiable raw-text correspondence (Appendix B). We do not claim linearized graph OT as novel (Kolouri et al., 2021; Beier et al., 2022); the novelty is the nested-ground-cost linearization, the shared cross-modal core, and the four-determinant characterization.

Future work A basepoint- and orientation-aware distortion bound for discrete measures would make the scope condition quantitative; Prop. 3 suggests the right object is the $L^2(\sigma)$ geometry of the entropic barycentric projection. Extending to Gromov–Wasserstein would remove the shared-atom assumption that makes the two orientations comparable. Untested instantiations include video (component index space as time), multimodal retrieval (Shaaban et al., 2025), and single-cell omics.

AI USE STATEMENT

(This section is required and does not count toward the page limit.) In this work, we used generative AI tools for [tasks with required disclosure]. We have not used generative AI tools for [other tasks with required disclosure], and [the rest of the required disclosure tasks] are not applicable to this work. Additionally, we used generative AI tools for [tasks with recommended disclosure]. We have reviewed all AI-assisted work. [Elaborate.] We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

(Recommended; does not count toward the page limit.) Proofs of all propositions are in Appendix A. Datasets, preprocessing, component-model fitting, reference construction, solver settings, seeds and hardware are specified in Appendix B, and complete numerical results in Appendix C. Anonymous source code will be provided as supplementary material.

REFERENCES

Florian Beier, Robert Beinert, and Gabriele Steidl. On a linear Gromov-Wasserstein distance. *IEEE Transactions on Image Processing*, 31:7292–7305, 2022. doi: 10.1109/TIP.2022.3221286.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pp. 2292–2300, 2013.

Alex Delalande and Quentin Mérigot. Quantitative stability of optimal transport maps under variations of the target measure. *Duke Mathematical Journal*, 172(17):3321–3357, 2023. doi: 10.1215/00127094-2022-0106.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl 1):5228–5235, 2004. doi: 10.1073/pnas.0307752101.

Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, 2022.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. doi: 10.1109/TBDATA.2019.2921572.

Varun Khurana, Harish Kannan, Alexander Cloninger, and Caroline Moosmüller. Supervised learning of sheared distributions using linearized optimal transport. *Sampling Theory, Signal Processing, and Data Analysis*, 21(1):Article 1, 2023. doi: 10.1007/s43670-022-00038-2.

Soheil Kolouri, Navid Naderializadeh, Gustavo K. Rohde, and Heiko Hoffmann. Wasserstein embedding for graph learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pp. 957–966. PMLR, 2015.

Yu A. Malkov and Dmitry A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836, 2020. doi: 10.1109/TPAMI.2018.2889473.

Facundo Mémoli. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, 11(4):417–487, 2011. doi: 10.1007/s10208-011-9093-5.

Quentin Mérigot, Alex Delalande, and Frédéric Chazal. Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3186–3196. PMLR, 2020.

Caroline Moosmüller and Alexander Cloninger. Linear optimal transport embedding: Provable Wasserstein classification for certain rigid transformations and perturbations. *Information and Inference: A Journal of the IMA*, 12(1):363–389, 2023. doi: 10.1093/imaiai/iaac023.

Christopher Morris, Nils M. Kriege, Franka Bause, Kristian Kersting, Petra Mutzel, and Marion Neumann. TuDataset: A collection of benchmark datasets for learning with graphs. *CoRR*, abs/2007.08663, 2020.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.

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

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

Bernhard Schmitzer. Stabilized sparse scaling algorithms for entropy regularized transport problems. *SIAM Journal on Scientific Computing*, 41(3):A1443–A1481, 2019. doi: 10.1137/16M1106018.

Mai A. Shaaban, Tausifa Jan Saleem, Vijay Ram Kumar Papineni, and Mohammad Yaqub. MOTOR: Multimodal optimal transport via grounded retrieval in medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention (MICCAI 2025)*, volume 15965 of *Lecture Notes in Computer Science*. Springer, 2025. doi: 10.1007/978-3-032-04978-0_44.

Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12:2539–2561, 2011.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1 (NeurIPS Datasets and Benchmarks 2021)*, 2021.

Matteo Togninalli, Elisabetta Ghisu, Felipe Llinares-López, Bastian Rieck, and Karsten Borgwardt. Wasserstein Weisfeiler-Lehman graph kernels. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 6436–6446, 2019.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6275–6284. PMLR, 2019.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, Heidelberg, 2009. ISBN 978-3-540-71049-3.

Wei Wang, Dejan Slepčev, Saurav Basu, John A. Ozolek, and Gustavo K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101(2):254–269, 2013. doi: 10.1007/s11263-012-0566-z.

Lingfei Wu, Ian En-Hsu Yen, Kun Xu, Fangli Xu, Avinash Balakrishnan, Pin-Yu Chen, Pradeep Ravikumar, and Michael J. Witbrock. Word mover's embedding: From Word2Vec to document embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4524–4534, Brussels, Belgium, 2018. Association for Computational Linguistics.

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin Solomon. Hierarchical optimal transport for document representation. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, pp. 1599–1609, 2019.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad Reza Namazi-Rad. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 3886–3893, 2022.

A Proofs

A.1 Proposition 1

With $p = 1$, $C^{(0)}_{jk} = \|A_j - A_k\|_2$ is a metric on $\mathcal{A}$. For a metric ground cost the Kantorovich functional is a metric on probability measures over $\mathcal{A}$ (Villani, 2009, Ch. 6): nonnegativity and symmetry are immediate; $W_1(\alpha, \beta) = 0$ forces $\alpha = \beta$ because the only admissible zero-cost plan is supported on the diagonal; and the triangle inequality follows by gluing, taking $\Pi^{\alpha\gamma}_{jl} = \sum_{k:\beta_k>0} \Pi^{\alpha\beta}_{jk} \Pi^{\beta\gamma}_{kl} / \beta_k \in \Gamma(\alpha, \gamma)$ and bounding its cost with the triangle inequality for $C^{(0)}$.

Hence $C^{(1)}_{st} = W_1(\bar{c}_s, \bar{c}_t; C^{(0)})$ inherits nonnegativity, symmetry and the triangle inequality, so it is a pseudometric; and $C^{(1)}_{st} = 0$ iff $\bar{c}_s = \bar{c}_t$, so it separates indices exactly when $s \mapsto \bar{c}_s$ is injective. Applying the first argument again with ground cost $C^{(1)}$ gives a metric on Δ_S when $C^{(1)}$ is a metric, and a pseudometric otherwise, with $W_1(\mu, \nu; C^{(1)}) = 0$ whenever μ and ν differ only by exchanging mass between components with identical atom distributions. For $p > 1$, $\|\cdot\|_2^p$ fails the triangle inequality, as three collinear points show. □

Remark 2. Component sparsification is applied *before* (1) is evaluated, so the proposition applies verbatim to sparsified components. Truncating an *object's* mixture at a threshold, as HOTT does at $1/(S+1)$, instead modifies the measures being compared and voids the third statement.

A.2 Proposition 2

Restricted to the supports of the two marginals, (2) has $n'm'$ variables and $n'+m'$ equality constraints, exactly one redundant because both marginals sum to one, so the constraint matrix has rank $n'+m'-1$ and a basic feasible solution has at most that many nonzeros. Every row i with $\mu_i > 0$ has at least one nonzero since it sums to μ_i. If r rows have two or more nonzeros then $\|\Pi\|_0 \geq (n'-r)+2r = n'+r$, so $r \leq m'-1$ and at least $n'-m'+1$ rows are singletons. For a singleton row i with its nonzero in column $s(i)$, $\Pi_{is(i)} = \mu_i$, so $\Phi_\sigma^{\text{SRC}}(\mu)_i = e_{s(i)}$. □

Remark 3. The bound does not depend on how concentrated μ is beyond its support cardinality, which is why the concentration hypothesis of §6.3 fails. With $n' = m' = S$ the counting argument alone forces only one singleton row; the 85 to 91% measured is an empirical property of the optimal vertex.

A.3 Proposition 3

The $\varepsilon \to 0^+$ statement is standard: the entropic objective converges to the linear one and cluster points of $\Pi^{\mu,\varepsilon}$ are optimal for (2) (Peyré & Cuturi, 2019, Ch. 4). For $\varepsilon \to \infty$, dividing the objective by ε shows minimizers converge to the maximizer of H over $\Gamma(\mu, \sigma)$, which is the independent coupling $\mu\sigma^\top$, unique because H is strictly concave. Substituting $\Pi_{is} = \mu_i\sigma_s$ into (3) gives $\Phi_\sigma^{\text{SRC}}(\mu)_i = \sum_s \sigma_s e_s$ and $\Phi_\sigma^{\text{REF}}(\mu)_i = \sum_s \mu_s e_s$, both constant in i. The first does not involve μ, so the source-indexed embedding becomes constant across objects and all pairwise distances vanish. The second is $E^\top\mu$, which retains first-order information about μ. □

B Protocol Details

Text Vocabulary pruning removes stopwords and words of fewer than three characters, Snowball stemming merges morphological variants with embeddings averaged over merged surface forms, and the vocabulary is restricted to GloVe-covered terms (300 dimensions). Topics are fitted with a collapsed Gibbs sampler at 1500 iterations for $S = 70$ topics, each represented by its 20 highest-weighted words and renormalized. Nested costs use $p = 1$ and are solved as linear programs. For BEIR the vocabulary is capped at 5000 terms and the sampler at 500 iterations; TREC-COVID is capped at 60,000 documents.

Graphs Table 5 gives per-dataset WL settings. The ten-feature structural descriptor is degree, log-degree, mean, max and min neighbour degree, core number, clustering coefficient, triangle count, WL depth, WL label frequency and neighbour-label entropy, standardized column-wise. On dense corpora (DD, IMDB-MULTI) clustering and triangles are $O(Vd^2)$ per graph and are replaced by min neighbour degree, giving a nine-feature $O(V+E)$ subset. Components are k-means clusters over atom coordinates with $S = 70$ and top-20-label sparsification.

Images Patch descriptors are 2×2 spatial cells by 8 orientation bins, magnitude-weighted and L_2-normalized, at patch size 7 and stride 3 (5 and 2 on USPS). The lowest-energy half of patches per image is suppressed. Descriptors are reduced to 16 dimensions by PCA and quantized to a 300-entry k-means codebook, with soft assignment over the three nearest codewords by inverse distance. Components are $S = 70$ clusters with top-20-atom sparsification, and corpora are subsampled to 200 images per class.

Reference, seeds and solvers The reference is built on the one-dimensional component index space with FWHM 10 unless stated. Component models and codebooks use seed 0; split r uses seed r. Exact plans use a network-simplex solver; entropic plans use a log-domain scaling iteration (Schmitzer, 2019) with ε expressed as a fraction of the largest cost entry, because $\exp(-C^{(1)}/\varepsilon)$ underflows otherwise (the largest entry is 9.50 on NFCORPUS), and convergence is verified by maximum marginal violation rather than by solver warnings. Text classification ran on an Intel i9-13900HX with 32 GB; all other measurements ran on cluster CPU nodes with 8 cores and 96 to 128 GB.

Raw-text correspondence for encoder-based component models Encoder-based component models require raw sentences, and reconstruction from a bag of words cannot recover sentence

structure, so we verified row-for-row correspondence between raw text and count matrices before use. Projecting each raw document onto the retained vocabulary and comparing against the count matrix by cosine similarity, four corpora verified: BBCSPORT (mean similarity 0.999, label agreement 1.000), CLASSIC (0.982, 1.000, recovered as a subset with four extra raw documents), AMAZON (0.934, 1.000) and OHSUMED (0.985, 1.000, under train-then-test concatenation). Two did not. TWITTER fails on label agreement 0.938, which its brevity explains: 9.7 average unique words per document over a vocabulary of 1205 terms after filtering makes many documents near-identical as count vectors. REUTERS has no concatenation of its split files reaching the required row count, and its staged text is pre-stemmed, so it would not test contextual encoding even if aligned. Both are excluded rather than run on an unverified pairing.

Table 5: Graph configurations. K_{full} is the WL dictionary size before truncation, K the retained support, d the atom coordinate dimension. Geometry is the coordinate regime: attr(d) is the attribute barycenter, struct the ten-feature descriptor and struct-9 the $O(V + E)$ subset used on dense corpora.

Dataset	n	cls.	r	K_{full}	K	d	geometry
MUTAG	188	2	2	214	174	10	struct
ENZYMES	600	6	2	10,650	500	18	attr(18)
PROTEINS	1113	2	1	300	297	1	attr(1)
PROTEINS_full	1113	2	1	300	297	29	attr(29)
DD	1178	2	1	253,231	300	9	struct-9
IMDB-MULTI	1500	3	2	—	300	9	struct-9
NCI1	4110	2	2	4387	300	10	struct
Mutagenicity	4337	2	2	4615	300	10	struct

Table 6: Text classification benchmarks.

Dataset	$\lvert D\rvert$	V	avg(w)	classes
BBCSPORT	737	3657	116.5	5
TWITTER	3108	1205	9.7	3
OHSUMED	9152	8261	59.4	10
CLASSIC	7093	5813	38.5	4
REUTERS	7674	5495	35.7	8
AMAZON	8000	16753	44.3	4

C Complete Numerical Results

Table 7: Text classification, mean test error across all reference counts and OT baselines. Bold marks the best linearized variant per corpus.

Dataset	LOTT-1	LOTT-2	LOTT-3	LOTT-4	LOTT-5	LOTT-10	LOTT-15	HOFTT	HOTT
BBCSPORT	0.164	0.139	0.132	0.098	0.079	**0.070**	0.081	0.041	0.036
TWITTER	0.342	0.338	0.326	**0.316**	0.331	0.321	0.326	0.342	0.350
OHSUMED	0.523	0.525	0.489	0.491	0.478	**0.462**	0.464	0.469	0.462
CLASSIC	0.082	0.066	0.055	0.062	0.057	**0.053**	0.079	0.057	0.051
REUTERS	0.113	0.105	0.106	0.097	0.103	**0.093**	0.096	0.058	0.062
AMAZON	0.133	0.113	0.125	0.115	0.114	**0.112**	0.114	0.112	0.110

Table 8: Text throughput normalized so exact pairwise HOTT= 1.0. WMD-20 truncates each document to its 20 most frequent words.

Dataset	WMD-20	HOFTT	LOTT-1	LOTT-5	LOTT-10	LOTT-15
BBCSPORT	0.374	0.307	11.467	12.683	12.524	12.593
TWITTER	0.397	0.290	98.579	101.807	101.073	101.646
OHSUMED	0.169	0.388	136.026	125.825	124.353	124.953
CLASSIC	0.275	0.212	93.333	87.735	86.156	84.983
REUTERS	0.149	0.177	92.032	91.417	89.976	89.441
AMAZON	0.074	0.287	182.299	168.049	168.388	165.111

Table 9: Encoder comparison. Test error and inference time, the latter normalized so LOTT-10 = 1.0; both embedding generation and k-NN classification are included, on a common CPU.

	SBERT		SBERT-large		DistilBERT		RoBERTa		BERT	
Dataset	err	time	err	time	err	time	err	time	err	time
BBCSPORT	0.054	40.9	0.027	245.5	0.034	90.9	0.095	193.2	0.014	181.8
TWITTER	0.278	2.4	0.268	15.9	0.305	7.1	0.292	14.0	0.291	14.1
OHSUMED	0.315	34.1	0.296	181.7	0.503	70.4	0.551	137.3	0.513	149.4
CLASSIC	0.038	20.3	0.021	113.2	0.037	46.2	0.058	86.8	0.039	93.0
REUTERS	0.067	24.3	0.059	148.2	0.086	62.4	0.091	112.6	0.084	137.4
AMAZON	0.056	42.6	0.049	238.0	0.121	91.9	0.131	164.5	0.078	179.9

Table 10: Left: the topic-proportion baseline, where raw component proportions are compared by Euclidean distance with no transport, throughput normalized to LOTT-10. Right: linear interpolation of LOTT-10 distances with BM25 scores.

	Topic proportions only		Lexical hybrid	
Dataset	test error	rel. thr.	LOTT-10	+BM25
BBCSPORT	0.065	1.656	0.070	**0.027**
TWITTER	0.336	0.422	**0.321**	0.323
OHSUMED	0.495	0.220	0.462	**0.395**
CLASSIC	0.058	0.172	0.053	**0.036**
REUTERS	0.072	0.260	0.093	**0.040**
AMAZON	0.131	0.215	0.112	**0.074**

Table 11: End-to-end wall-clock in seconds. Setup (a, b) is the one-time component fit and nested cost matrix; inference (c, d) is per-corpus embedding and k-NN query, averaged over the tested FWHM values.

Dataset	(a) prep/fit	(b) nested cost	(c) embed	(d) query	total
TWITTER	73.76	0.78	6.30	0.93	81.77
BBCSPORT	174.00	0.98	1.33	0.16	176.47
REUTERS	551.76	1.37	13.96	2.94	570.04
CLASSIC	566.40	1.39	12.45	2.48	582.73
OHSUMED	1073.11	2.21	16.85	3.95	1096.13
AMAZON	1633.53	6.47	14.28	3.18	1657.46

Table 12: Reference-width sensitivity on the two corpora with the largest label spaces. Times are in seconds; throughput is test-train pairs per second. Accuracy responds materially while throughput is invariant.

Dataset	FWHM	test err.	embed	fit	pred.	throughput
OHSUMED	10	0.5398	16.87	1.14	3.53	728,842
	50	0.4795	17.04	1.19	2.84	745,101
	100	0.4646	16.81	0.79	2.88	766,704
	500	0.4563	16.83	0.79	2.92	764,945
	1000	**0.4471**	16.71	0.77	2.94	769,131
REUTERS	10	0.0990	13.87	0.74	2.20	657,076
	50	**0.0891**	13.97	0.76	2.00	660,265
	100	0.0902	13.82	0.72	2.35	654,051
	500	0.0954	14.24	0.77	2.54	629,189
	1000	0.0933	13.89	0.68	1.98	667,250

Table 13: Curated retrieval, all arms and cutoffs. NFCORPUS evaluates 257 of 323 queries after 66 are emptied by vocabulary filtering; SCIFACT and TREC-COVID drop none. TREC-COVID is capped at 59,186 documents and has only 50 queries, so its ordering should be read as indicative.

Corpus	Method	NDCG@10	MAP	P@10	MRR	q/s
NFCORPUS	HOTT (exact)	0.1121	0.1836	0.1121	0.2104	1.51
	LOTT-1	0.0418	0.0892	0.0416	0.1048	548.9
	LOTT-5	0.0792	0.1509	0.0782	0.1843	493.4
	LOTT-10	0.0855	0.1554	0.0837	0.1802	502.6
	BM25	0.3231	0.4394	0.2572	0.5135	8604.6
	Sentence encoder	0.3388	0.4673	0.2708	0.5607	1670.9
SCIFACT	HOTT (exact)	0.1533	0.1286	0.0307	0.1300	1.06
	LOTT-1	0.0558	0.0465	0.0110	0.0477	559.2
	LOTT-5	0.1006	0.0849	0.0210	0.0854	499.7
	LOTT-10	0.1239	0.1036	0.0250	0.1048	487.8
	BM25	0.5454	0.5088	0.0750	0.5149	760.6
	Sentence encoder	0.6393	0.5973	0.0873	0.6013	831.2
TREC-COVID	HOTT (exact)	0.0428	0.1753	0.0960	0.1972	0.09
	LOTT-1	0.0301	0.1196	0.0500	0.1580	525.0
	LOTT-5	0.0550	0.1605	0.0860	0.1863	436.8
	LOTT-10	0.0491	0.1663	0.0880	0.2024	459.2
	BM25	0.2512	0.6877	0.5960	0.8292	119.8
	Sentence encoder	0.3040	0.4894	0.3500	0.6779	611.8

Table 14: Entropic sweep on NFCORPUS, source-indexed LOTT-10. Deltas and t are paired against the exact-plan value of 0.0862; exact pairwise HOTT attains 0.1121. Marginal violation is the maximum over both marginals, confirming the sweep measures regularization rather than solver failure.

ε	NDCG@10	Δ vs. exact plan	t	nz/row	marg. viol.
0.05	0.0914	$+0.0053$	$+1.36$	33.60	7.8×10^{-10}
0.10	0.0909	$+0.0047$	$+1.12$	69.09	1.5×10^{-10}
0.15	0.0985	$+0.0123$	$+2.49$	70.00	1.1×10^{-10}
0.20	**0.0996**	$+0.0134$	$+2.09$	70.00	1.4×10^{-10}
0.25	0.0959	$+0.0097$	$+1.33$	70.00	6.7×10^{-14}
0.30	0.0885	$+0.0024$	$+0.30$	70.00	1.3×10^{-15}
0.40	0.0768	-0.0094	-1.17	70.00	2.2×10^{-16}

Table 15: Graph 2×2: test error (mean$\pm$std over 10 splits) and throughput relative to each level's own pairwise baseline. Source-indexed, $k = 10$. IMDB-MULTI is degenerate under both orientations (Table 19).

Dataset	WWL	L-WWL-10	H-WWL	L-H-WWL-10	flat	hier
MUTAG	$0.102 \pm .039$	$0.249 \pm .089$	$0.109 \pm .035$	$0.200 \pm .066$	$24\times$	$52\times$
ENZYMES	$0.687 \pm .026$	$0.689 \pm .054$	$0.689 \pm .027$	$\mathbf{0.683} \pm .029$	$33\times$	$118\times$
PROTEINS	$0.296 \pm .018$	$0.307 \pm .026$	$0.301 \pm .021$	$\mathbf{0.296} \pm .017$	$93\times$	$291\times$
PROTEINS_full	$0.302 \pm .022$	$0.312 \pm .022$	$0.306 \pm .027$	$\mathbf{0.301} \pm .023$	$72\times$	$238\times$
DD	$0.394 \pm .016$	$0.454 \pm .038$	$0.391 \pm .019$	$0.440 \pm .028$	$88\times$	$204\times$
IMDB-MULTI	$0.578 \pm .029$	$0.593 \pm .035$	$0.570 \pm .032$	$0.598 \pm .033$	$108\times$	$308\times$
NCI1	$0.244 \pm .012$	$\mathbf{0.219} \pm .013$	$0.256 \pm .014$	$0.261 \pm .013$	$325\times$	$1115\times$

Table 16: Paired hierarchical minus flat difference in test error on identical splits; negative favours the nested cost. Bold marks $|t| \geq 2$.

Dataset	k=1	k=5	k=10
MUTAG	$+0.102$ ($+3.3$)	$\mathbf{-0.040}$ (-2.6)	$\mathbf{-0.049}$ (-3.2)
ENZYMES	$+0.003$ ($+0.1$)	$+0.011$ ($+0.6$)	-0.005 (-0.3)
PROTEINS	$\mathbf{-0.036}$ (-3.1)	-0.003 (-0.6)	-0.011 (-1.3)
PROTEINS_full	$+0.009$ ($+0.9$)	-0.008 (-1.1)	-0.012 (-1.5)
DD	$\mathbf{-0.043}$ (-3.7)	$\mathbf{-0.020}$ (-3.1)	-0.014 (-1.1)
IMDB-MULTI	$+0.001$ ($+0.1$)	$+0.004$ ($+0.3$)	$+0.006$ ($+0.5$)
NCI1	$\mathbf{+0.036}$ ($+7.5$)	$\mathbf{+0.033}$ ($+7.2$)	$\mathbf{+0.042}$ ($+8.2$)

Table 17: Graph retrieval, five repeats, same-class relevance. Query-time throughput is queries divided by query embedding plus search time; the paired ΔMAP is against the arm's own exact pairwise baseline on shared splits.

Dataset	Method	P@1	MAP	MRR	R-prec.	q/s	ΔMAP (t)
NCI1	WWL	0.7747	0.7506	0.8553	0.6695	0.8	—
	L-WWL-10	**0.7927**	**0.7616**	**0.8647**	**0.6767**	333.7	$+0.0111$ ($+4.1$)
	H-WWL	0.7499	0.7376	0.8401	0.6596	0.8	—
	L-H-WWL-10	0.7533	0.7248	0.8460	0.6339	1169.4	-0.0128 (-4.3)
Mutagenicity	WWL	0.7539	0.7454	0.8458	**0.6736**	0.8	—
	L-WWL-10	**0.7664**	**0.7493**	**0.8539**	0.6705	357.9	$+0.0038$ ($+1.2$)
	H-WWL	0.7415	0.7376	0.8374	0.6690	0.8	—
	L-H-WWL-10	0.7433	0.7293	0.8405	0.6524	1255.8	-0.0083 (-3.2)
ENZYMES	WWL	0.4167	0.4051	0.5684	0.2158	2.8	—
	L-WWL-10	**0.4233**	**0.4380**	**0.5783**	0.2085	123.5	$+0.0329$ ($+3.9$)
	H-WWL	0.3767	0.3932	0.5275	0.2105	5.8	—
	L-H-WWL-10	0.3667	0.3977	0.5290	**0.2317**	1127.2	$+0.0045$ ($+0.5$)
PROTEINS_full	WWL	0.6811	**0.7008**	0.7997	**0.6344**	3.1	—
	L-WWL-10	**0.6847**	0.6907	0.7989	0.6187	328.4	-0.0101 (-1.8)
	H-WWL	0.6937	0.7001	**0.8013**	0.6264	3.2	—
	L-H-WWL-10	0.6685	0.6890	0.7939	0.6281	1214.6	-0.0111 (-1.4)
DD	WWL	**0.5847**	0.6128	**0.7369**	0.5466	2.0	—
	L-WWL-10	0.5525	0.6087	0.7209	0.5242	246.5	-0.0041 (-0.4)
	H-WWL	0.5814	**0.6155**	0.7323	**0.5519**	2.6	—
	L-H-WWL-10	0.5458	0.6009	0.7203	0.5325	1100.2	-0.0146 (-3.2)

Table 18: Image classification (test error, ten repeats) and MNIST retrieval, source-indexed. Throughput is pairs per second for classification and queries per second for retrieval.

	Flat			Hierarchical		
Dataset	exact	lin.-1	lin.-10	exact	lin.-1	lin.-10
MNIST	0.1478	0.3118	0.4258	0.2018	0.3194	0.3278
F-MNIST	0.2640	0.3128	0.3254	0.3090	0.3262	0.3388
USPS	0.0956	0.2238	0.2496	0.1186	0.2424	0.2402

MNIST retrieval: P@1 / MAP / q·s^{-1}
exact flat 0.821 / 0.778 / 0.5 lin. flat 0.759 / 0.715 / 158.8
exact hier 0.743 / 0.712 / 1.4 lin. hier 0.627 / 0.611 / 726.8

Table 19: All 24 orientation arms, ordered by linearization gap. *blocks* is the median number of distinct barycentric blocks under source and reference indexing, out of the total shown. Dashes mark arms where Φ^{BOTH} and the validation selector were not evaluated.

Arm	mod.	k	exact	Φ^{SRC}	Φ^{REF}	Φ^{BOTH}	sel.	gap	Δ	t
MNIST/flat	image	10	0.1478	0.4258	0.1910	0.1840	0.1894	+0.278	−0.235	−39.9
MNIST/flat	image	5	0.1478	0.3966	0.2156	0.1838	0.1866	+0.249	−0.181	−34.3
MNIST/flat	image	1	0.1478	0.3118	0.2426	0.1864	0.1864	+0.164	−0.069	−9.7
USPS/flat	image	10	0.0956	0.2496	0.1088	0.1050	0.1062	+0.154	−0.141	−51.5
MUTAG/flat	graph	10	0.1021	0.2489	0.1468	—	—	+0.147	−0.102	−6.0
MNIST/hier	image	10	0.2018	0.3278	0.2382	0.2444	0.2388	+0.126	−0.090	−15.0
MNIST/hier	image	5	0.2018	0.3264	0.2446	0.2390	0.2394	+0.125	−0.082	−11.9
USPS/hier	image	10	0.1186	0.2402	0.1452	0.1376	0.1396	+0.122	−0.095	−17.2
MNIST/hier	image	1	0.2018	0.3194	0.3118	0.2572	0.2572	+0.118	−0.008	−1.4
MUTAG/hier	graph	10	0.1085	0.2000	0.1213	—	—	+0.092	−0.079	−5.5
F-MNIST/flat	image	10	0.2640	0.3254	0.2886	0.2614	0.2652	+0.061	−0.037	−11.0
DD/flat	graph	10	0.3939	0.4542	0.3959	—	—	+0.060	−0.058	−3.9
DD/hier	graph	10	0.3905	0.4400	0.3824	—	—	+0.050	−0.058	−6.2
F-MNIST/hier	image	10	0.3090	0.3388	0.3156	0.2958	0.2990	+0.030	−0.023	−4.1
IMDB-MULTI/hier	graph	10	0.5699	0.5984	0.5939	0.5891	0.5995	+0.029	−0.005	−0.7
IMDB-MULTI/flat	graph	10	0.5781	0.5928	0.5779	0.5784	0.5829	+0.015	−0.015	−2.7
PROTEINS/flat	graph	10	0.2957	0.3072	0.3358	0.2982	0.3090	+0.012	+0.029	+2.8
PROTEINS_full/flat	graph	10	0.3022	0.3122	0.3086	—	—	+0.010	−0.004	−0.3
NCI1/hier	graph	10	0.2556	0.2608	0.2772	—	—	+0.005	+0.016	+3.0
ENZYMES/flat	graph	10	0.6873	0.6887	0.7460	—	—	+0.001	+0.057	+2.7
PROTEINS/hier	graph	10	0.3014	0.2964	0.3147	0.2993	0.2978	−0.005	+0.018	+2.6
PROTEINS_full/hier	graph	10	0.3061	0.3007	0.3391	—	—	−0.005	+0.038	+5.8
ENZYMES/hier	graph	10	0.6893	0.6833	0.7613	—	—	−0.006	+0.078	+7.0
NCI1/flat	graph	10	0.2437	0.2193	0.2785	—	—	−0.024	+0.059	+10.5

Table 20: NCI1 support-truncation sweep at fixed $S = 70$. The hierarchical penalty grows with the atom-set size while the flat arm is flat in K, isolating coarsening rather than truncation.

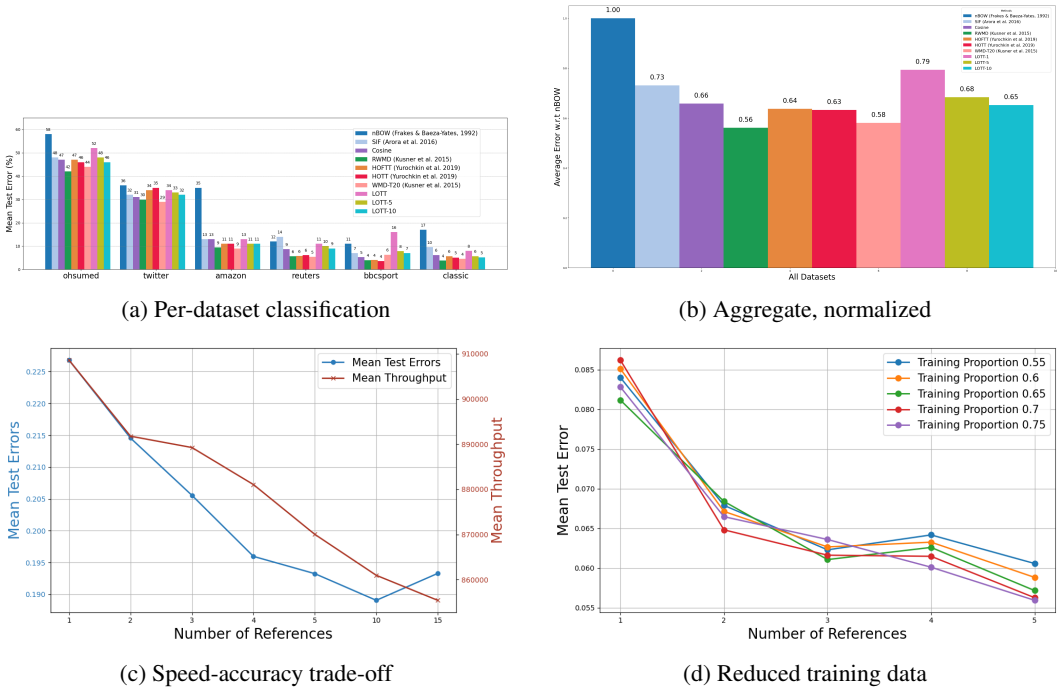
K	flat dim.	L-WWL-10	L-H-WWL-10	WL-subtree	hier − flat (t)
300	3000	**0.2193**	0.2608	0.2430	+0.042 (+8.2)
1000	10000	0.2366	0.2905	**0.2075**	+0.054 (+15.4)
3000	30000	0.2228	0.3255	**0.2086**	+0.103 (+20.3)

Table 21: Component models under all three orientations at matched $S = 70$. Image generative rows use soft codeword assignment, which disadvantages a model requiring integral counts: on USPS a hard-assignment control raises retained mass from 8.5 to 18.0 counts per image. Text corpora use $k = 10$ on the canonical split, so no paired variance is available; BERTopic returns an emergent component count, 6 on BBCSPORT, and failed on AMAZON with a non-finite class-based weighting matrix.

		exact			Φ^{REF}			Φ^{SRC}		
Dataset	mod.	clus.	gen.	obj.	clus.	gen.	obj.	clus.	gen.	obj.
MUTAG	graph	0.109	0.181	0.109	0.121	0.170	0.121	0.200	0.200	0.349
ENZYMES	graph	0.689	0.643	0.699	0.761	0.735	0.700	0.683	0.723	0.705
PROTEINS_full	graph	0.306	0.307	0.313	0.339	0.346	0.317	0.301	0.344	0.310
DD	graph	0.391	0.434	0.404	0.382	0.428	0.412	0.440	0.429	0.399
NCI1	graph	0.256	0.230	0.228	0.277	0.285	0.242	0.261	0.311	0.262
USPS	image	0.119	0.299	0.100	0.145	0.349	0.120	0.240	0.364	0.117
F-MNIST	image	0.309	0.354	0.335	0.316	0.391	0.332	0.339	0.419	0.340
MNIST	image	—	0.265	—	—	0.370	—	—	0.394	—
BBCSPORT	text	—	0.043	0.081	—	0.032	0.070	—	0.060	0.065
AMAZON	text	—	—	—	—	0.093	0.077	—	0.105	0.081

D SUPPORTING FIGURES

Figure 5 collects the text-instantiation panels referenced in §6.1: per-dataset classification error for all arms; the aggregate normalized by a bag-of-words baseline, where values below one indicate improvement over that baseline; the speed-accuracy trade-off as reference components are added, with error falling monotonically while throughput declines gradually; and robustness under reduced supervision on CLASSIC, where the three-component configuration is stable across training proportions from 0.55 to 0.75. The component-count trend in panel (c) is the text counterpart of the image trend whose apparent inversion is resolved by the orientation correction of §6.2.



(a) Per-dataset classification

(b) Aggregate, normalized

(c) Speed-accuracy trade-off

(d) Reduced training data

Figure 5: Text instantiation, supporting panels for §6.1.

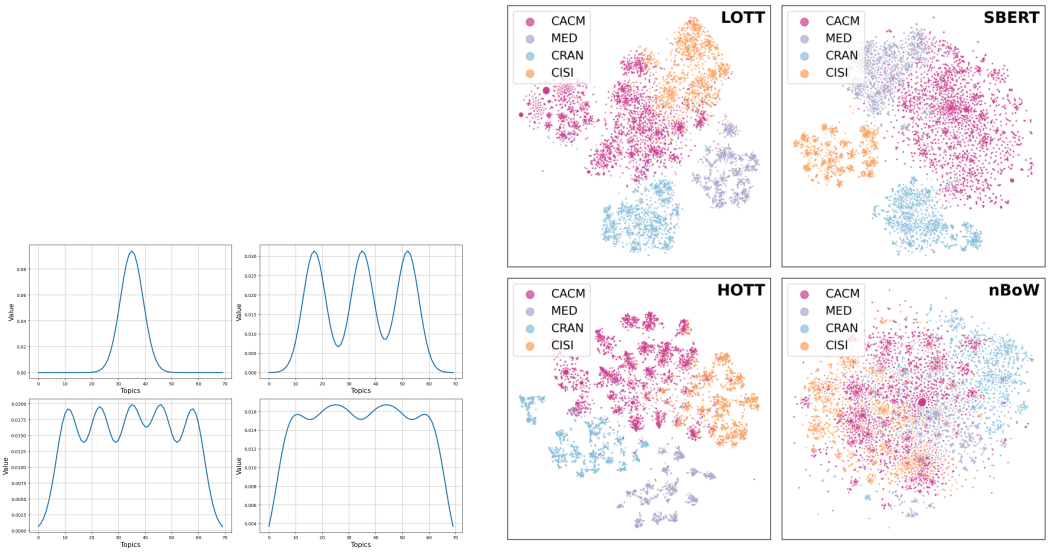


Figure 6: Left: the reference σ over the 70 component indices for one, two, five and ten components at fixed FWHM. Right: t-SNE of CLASSIC under a normalized bag-of-words baseline, the linearized transport embedding, exact hierarchical transport, and a sentence encoder.

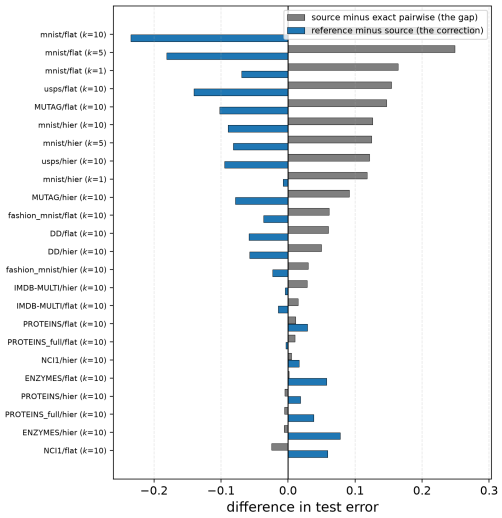


Figure 7: Per-arm decomposition underlying Fig. 1: the linearization gap beside the correction from reference-indexing, sorted by gap.

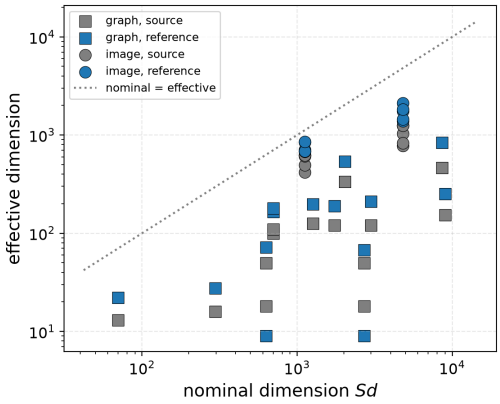


Figure 8: Effective against nominal embedding dimension. Source-indexing wastes most of the nominal dimension on sparse measures.

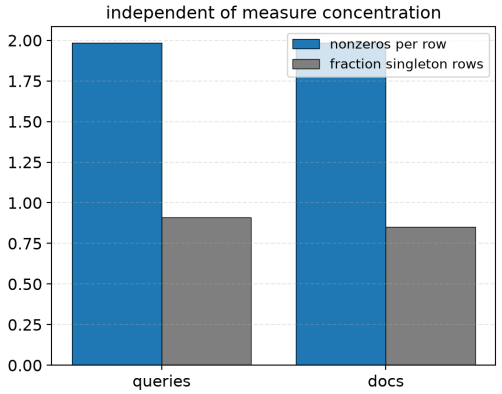


Figure 9: Nonzeros per row and singleton fraction for documents and queries separately, both independent of measure concentration, against the basic-solution bound $n' + m' - 1$, which is attained exactly.

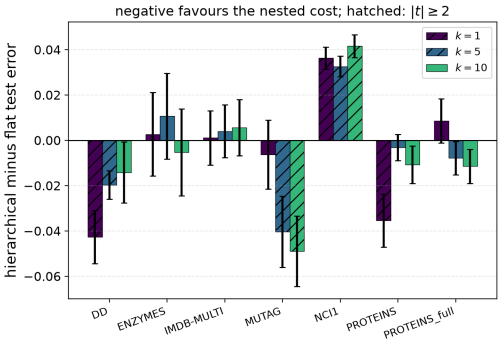


Figure 10: Paired hierarchical minus flat difference on identical splits. Negative favours the nested cost; hatching marks $|t| \geq 2$.

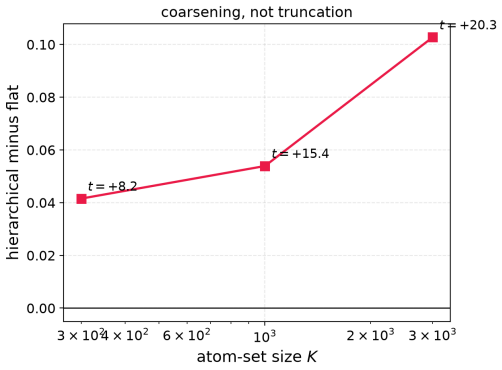


Figure 11: The hierarchical penalty grows with atom-set size at fixed S, isolating coarsening rather than truncation.

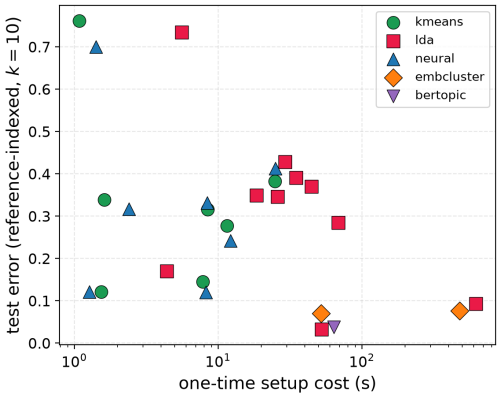


Figure 12: Accuracy against one-time setup cost. The generative component model is both slower to fit and usually less accurate.

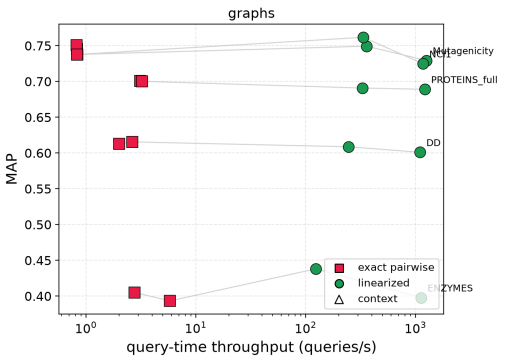


Figure 13: Graph retrieval quality against query-time throughput. Linearized arms sit three orders of magnitude to the right at comparable quality.

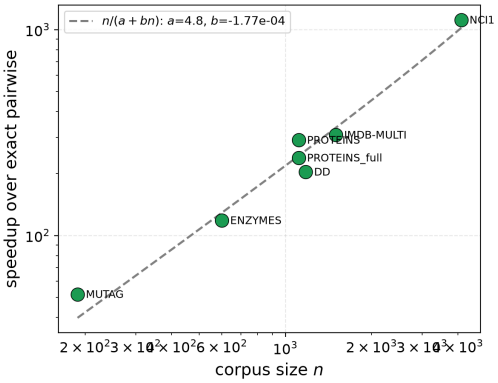


Figure 14: Speedup against corpus size with the fit $n/(a+bn)$ overlaid, and the n-independence of exact pairwise throughput that underwrites it.

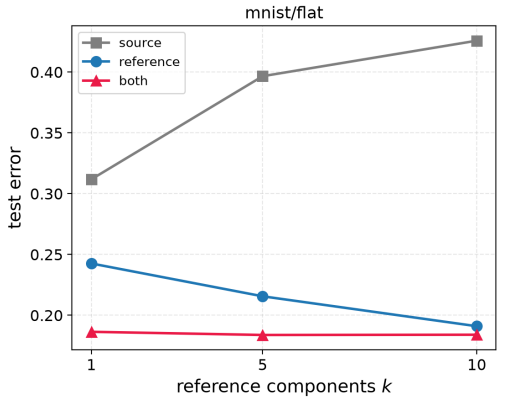


Figure 15: Reference component count under each orientation. The trend has opposite sign in the two conventions.

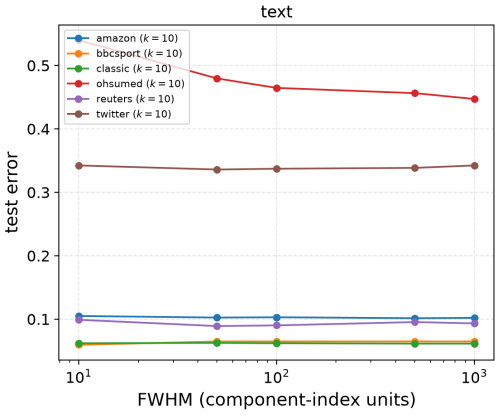
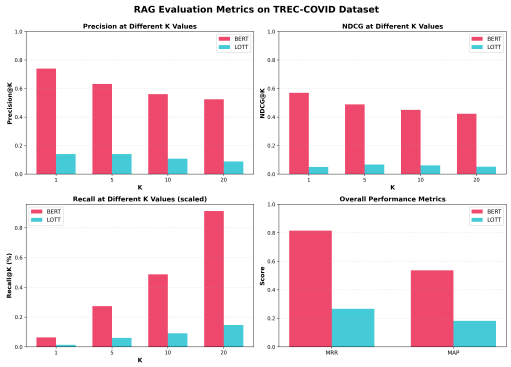


Figure 16: Test error against reference width on text, one line per corpus and component count; throughput is invariant, so the accuracy benefit of a wider reference is free.

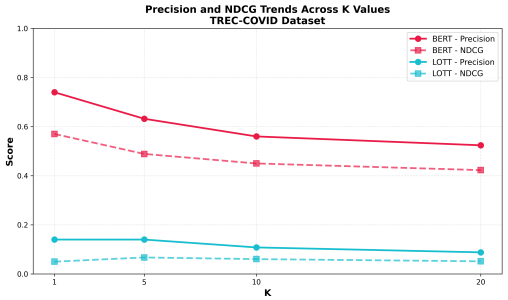
E Retrieval-Augmented Generation Pipelines

Beyond the controlled comparison of §6.3, we evaluated the text instantiation inside a two-stage retrieval pipeline on TREC-COVID, a biomedical corpus whose diverse topic distributions play to the strengths of a transport formulation. This setting is of practical interest because retrieval latency determines the responsiveness of a retrieval-augmented generation system, and because the transport formulation offers interpretability that purely learned embeddings lack: a similarity can be traced back to component proportions and to the geometric relationships the embedding preserves.

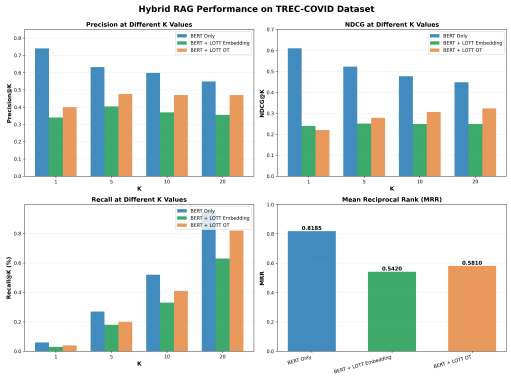
In the single-stage configuration (Fig. 17a, b) we built separate approximate nearest-neighbour indices (Johnson et al., 2019) from transport-based and encoder-based embeddings and computed precision, recall, NDCG and MRR at several cutoffs; encoder-based retrieval was stronger. In the two-stage configuration (Fig. 17c, d), following Shaaban et al. (2025), a larger candidate pool is retrieved and then reranked; direct encoder retrieval again outperformed the pipeline in which the transport embedding performed the reranking. We include these results because the hybrid framing remains the most promising practical use of transport geometry in retrieval and because the ordering they establish motivated the controlled comparison of §6.3, which holds the geometry fixed and varies only the solver and the orientation, and which we regard as the methodologically cleaner evidence. The complementarity established in Table 10 suggests the productive configuration is fusion rather than substitution.



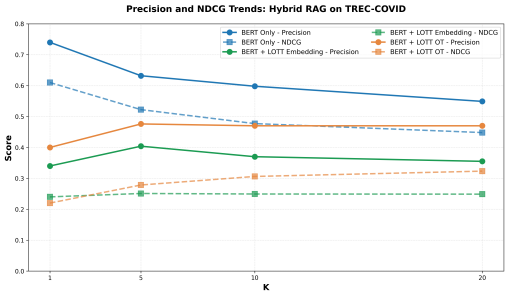
(a) Single-stage retrieval



(b) Cutoff trends



(c) Two-stage pipeline



(d) Two-stage cutoff trends

Figure 17: Two-stage retrieval pipelines on TREC-COVID, superseded for the purposes of this paper by the controlled comparison in §6.3.

F Negative and Inconclusive Results

Measure concentration does not explain the retrieval gap Effective support, measured as exp of the entropy of the component mixture, is 3.72 for queries (tenth percentile 1.23), 9.94 for documents (6.35) and 69.90 for the reference, out of 70 atoms. Per query, the Spearman correlation between effective support and the exact-minus-linearized NDCG@10 gap is -0.016 ($p = 0.80$), and the gap is

flat across concentration terciles at $+0.0246$, $+0.0308$ and $+0.0226$ from most to least concentrated, while exact transport itself degrades from 0.148 to 0.086 across the same terciles. Mean top-ten overlap between the exact and linearized rankings is only 0.310, so the two methods are differently rather than uniformly wrong. The transport plan is equally degenerate on both sides, with 1.99 nonzeros per row and a singleton fraction of 0.895 for queries against 0.846 for documents, exactly as Prop. 2 implies.

Metric diffusion is not an effective remedy Smoothing the query measure along the learned component metric, $\tilde{\mu} \propto \mu K_\tau$ with $K_\tau = \exp(-C^{(1)}/\tau)$ row-normalized, produces a clean inverted-U in τ, peaking at $+0.0034$ NDCG@10 at $\tau = 0.25\bar{C}$ where effective support rises from 3.7 to 33.9, then falling to -0.0434 at $\tau = \bar{C}$. The peak is within noise at this query count. Acting on the plan helps; acting on the measure does not.

Metric quality and linearizability are different objectives On ENZYMES the generative component model yields a better exact pairwise metric (0.643 against 0.689 for clustering) but a worse linearization (0.723 against 0.683 at source indexing). A component model producing a sharper metric is not necessarily easier to linearize, which cautions against selecting the component model on pairwise performance alone.