SEMANTIC MINHASH: BRINGING MINHASH TO EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

MinHash and locality-sensitive hashing (LSH) provide reusable signatures and bucket-based candidate generation for set similarity, but classical set MinHash does not directly capture the semantic geometry of dense embeddings. We present Semantic MinHash (SMinHash), which constructs a distance-annotated set for each embedding using shared sampled references and takes the minimum-distance index directly as its hash value. Repeated selections produce a discrete signature that brings the MinHash processing pattern to embedding retrieval, similarity graphs and duplicate screening. Building on nearest-reference winner-take-all selection, we derive an exact finite-reference collision law with shared-order ties, conditional locality and candidate-precision guarantees, and a maximum-load bound based on reference coverage and concentration. Experiments on text, image and multimodal embeddings demonstrate selective candidate generation: data references reduce mean candidate fractions on eight datasets against an equally tuned spherical-reference control, while semantic screening and complete-task comparisons establish practical benefits at the evaluated operating points. These results connect embedding-to-set construction, collision analysis and empirical task performance, providing a common basis for applying minimum-index signatures and LSH to semantic embeddings.

1 INTRODUCTION

MinHash-based locality-sensitive hashing (LSH) offers a reusable approach to similarity processing: repeated minimum selections produce compact signatures, matching signature bands form buckets, and original objects verify the resulting candidates (Broder, 1997; Lv et al., 2007). The same representation supports retrieval, similarity graphs and duplicate detection. Once the random functions are fixed, objects can be encoded independently and signatures reused across tasks, making this approach attractive for large collections with repeated similarity workloads.

Semantic embeddings represent text, images and other objects through their geometric relationships. Classical set MinHash, however, operates on discrete elements and measures set overlap. Treating a dense vector's nonzero coordinates as a set discards the values that distinguish embeddings, while equality of floating-point entries does not express their proximity. Weighted MinHash addresses weighted Jaccard similarity (Ioffe, 2010). Our objective is to bring the minimum-selection and LSH-bucketing approach to embedding geometry by constructing suitable sets from the vectors.

We present *Semantic MinHash* (SMinHash). Each round samples an ordered set of ψ reference embeddings shared by all inputs. For an embedding x, its distances to these references define a set of reference-index–distance pairs. The index attaining the minimum distance is the hash value, with ties resolved by the shared order. Repeating this construction independently T times yields $H_T(x) = (h_1(x), \ldots, h_T(x))$. These indices directly support banded LSH or an agreement-count threshold, followed by original-vector verification. Reference sets sampled from the data adapt the representation to the embedding distribution.

This construction makes the link to set MinHash explicit. For nonempty set indicators and a full unit-vector reference dictionary, it recovers the classical MinHash collision probability. More generally, it is winner-take-all (WTA) selection on negative reference distances (Yagnik et al., 2011), connecting established minimum-selection and nearest-reference ideas (Kang & Jung, 2012; Qin et al., 2019; Xu & Ting, 2025; Zhang et al., 2026) to a common embedding signature.

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.

The theory connects the constructed sets to both similarity and computational work. We derive an exact collision law for a finite shared reference pool, including distance ties, and propagate it through banded and collision-counting LSH. Under explicit geometric and pair-distribution conditions, the resulting guarantees describe when nearby embeddings collide, distant embeddings separate, and admitted pairs are likely to be close. A maximum-load bound then relates reference coverage and reverse-neighbor concentration to fitted bucket populations. Together, these results identify how reference selection and signature parameters govern candidate quality and workload.

Experiments test these properties across retrieval, semantic duplicate screening, graphs and complete threshold joins. Against equally tuned spherical references, data-derived sets reduce mean candidate fractions on all eight datasets and mean query time on seven. Semantic screening retains most cosine-positive pairs while reducing verification volume. Complete joins on COCO and Deep improve construction-plus-join cost over GPU-assisted IVF-Flat at the reported operating points. Graph studies and scaling to 9.99 million vectors further examine reuse of the signature mechanism.

We summarize our contributions as follows:

- We formulate an embedding-to-set construction whose minimum-distance indices directly support MinHash-style signatures and LSH candidate generation, with an explicit connection to classical set MinHash.
- We establish an exact finite-reference collision law, conditional locality and candidate-precision guarantees, and a maximum-load bound that links reference coverage and concentration to candidate work.
- We evaluate reference adaptation and complete similarity tasks across diverse embeddings, demonstrating selective retrieval, semantic screening, and graph and join processing with reusable signatures.

2 SEMANTIC MINHASH

SMinHash constructs a discrete signature directly from distances to shared reference sets. Each set supplies one minimum-distance index. The same signature supports banded LSH and collision-counting LSH; these differ only in how coordinate agreements produce candidates. Figure 1 illustrates the original banded construction. Primary experiments use $T = 100$; larger signatures and query multiprobing are development extensions.

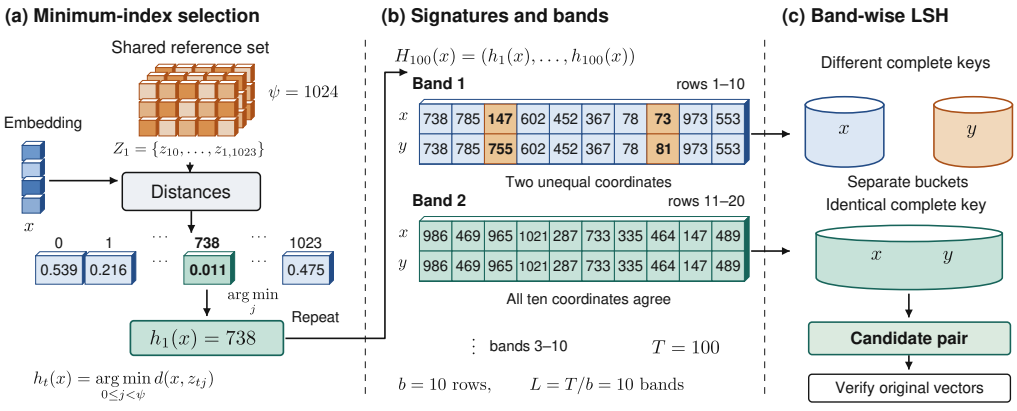

Figure 1: **Minimum-distance indices are the hash signature.** Each round samples a shared reference set and returns the index of its minimum distance. The resulting T indices are grouped directly into bands; a matching band produces a candidate for vector verification. The illustrative example uses $T = 100$, $\psi = 1024$, ten bands and ten rows per band. Selected distances from the first round have minimum index 738. Band 1 differs at two coordinates; band 2 matches completely and admits the pair. Omitted distances and bands are marked by ellipses.

From an embedding to minimum indices. Let $D = (z_1, \ldots, z_n)$ be a fixed reference collection in a metric space. Fix $2 \le \psi \le n$. Independently for each round $t \in [T]$, sample ψ reference items uniformly without replacement and in a uniform order, $Z_t = (z_{t0}, \ldots, z_{t,\psi-1})$. Database and query inputs share these ordered sets. For an embedding x, the distance-annotated set $A_t(x) = \{(j, d(x, z_{tj})) : j \in \{0, \ldots, \psi - 1\}\}$ determines

$$h_t(x) = \operatorname*{arg\,min}_{j \in \{0,\ldots,\psi-1\}} d(x, z_{tj}), \qquad H_T(x) = (h_1(x), \ldots, h_T(x)). \tag{1}$$

Ties select the smallest local index. Thus $H_T(x)$ contains exactly T minimizing indices in $\{0, \ldots, \psi - 1\}$, not distances. The equivalent set $\{(t, h_t(x)) : t \in [T]\}$ distinguishes indices from different reference sets by their round identifiers.

This completes the signature; no second MinHash is applied. Equivalently, it is WTA hashing of the reference features $s_D(x)_i = -d(x, z_i)$ under the same order and ties (Yagnik et al., 2011). Only sampled distances need evaluation. For cosine tasks, normalized vectors give the same nearest-reference ranking under cosine and Euclidean distance.

The connection to classical MinHash is exact in a special case: encode nonempty sets $A \subseteq [n]$ as indicator vectors and choose the reference dictionary $D = (e_1, \ldots, e_n)$ of unit vectors. With $\psi = n$, the first nearest reference is the first member of A in a random permutation, giving collision probability $J(A, B)$ (Broder, 1997). This is WTA's known binary specialization (Yagnik et al., 2011, Sec. 4.4). Partial sampling can add collisions when no set member is sampled; Appendix A.2 derives the correction. For general embeddings and reference collections, the collision quantity is nearest-reference agreement (Section 3).

Two candidate rules for one signature. For the original band-OR rule, choose b dividing T, giving $L = T/b$ disjoint bands with keys

$$B_\ell(x) = \big(\ell, h_{(\ell-1)b+1}(x), \ldots, h_{\ell b}(x)\big). \tag{2}$$

A pair is a candidate when any key agrees. Collision-counting LSH instead requires at least $m \in \{1, \ldots, T\}$ agreeing coordinates, regardless of their positions:

$$S_T(x, y) = \sum_{t=1}^{T} \mathbf{1}\{h_t(x) = h_t(y)\}, \qquad A_m(x, y) = \{S_T(x, y) \ge m\}. \tag{3}$$

Counting changes the candidate rule, not the encoder. It is an established LSH selection principle; Appendix A.11 provides its context and proofs for these signatures.

The count predicate can be evaluated by scanning packed signatures or by a lossless bucket prefilter. Partition the coordinates into $L_c = T - m + 1$ disjoint, nonempty, balanced groups and index their subtuples separately. A passing pair has at most $T - m$ mismatches, so at least one group must match completely. Unioning these buckets and verifying the original T coordinates therefore recovers exactly A_m. Constant padding of shorter keys is excluded from the count. Full-key checks remove fingerprint aliases in either bucket index.

Vector verification and execution. Original-vector verification ranks retrieval candidates, produces graph edges, or tests join and duplicate-screening thresholds. Self-queries reuse database signatures. Shared reference IDs permit each distinct distance to be computed once and gathered into its original round-local position, preserving the hash in exact arithmetic; numerical near ties can affect computed indices. Interval refinement reduces FP64 candidate scoring while preserving full-candidate top-k and ID tie-breaking. Appendix F.3 gives the refinement, numerical guards, and memory costs.

With $U \le \min(n, T\psi)$ distinct sampled references, encoding N database vectors and Q queries costs $O((N + Q)(Ud + T\psi))$, versus $O((N + Q)T\psi d)$ without reuse. Database signatures store NT indices, with $O(Ud + T\psi)$ shared reference and map storage. An index with $L_* = L$ or L_c groups forms keys in $O(NT)$ work and sorts postings in $O(L_* N \log N)$; hash tables allow expected $O(NL_*)$ insertions after key formation. Count verification and vector scoring remain data dependent: a full count scan costs $O(QNT)$, and all candidates may require $O(QNd)$ distance work. Our static refinement workspace uses $O(QN)$ memory, even for sparse candidates. These costs do not guarantee sublinear search or linear-time graph construction;

Reusing signatures across tasks. In exact arithmetic, fixed ordered reference sets make encoding independent of other inputs in the collection. Appended items can therefore receive compatible signatures without changing existing codes. For fixed T, changing a valid band width or agreement threshold also reuses the stored indices without evaluating reference distances again. Each task still requires its own operating-point selection and original-vector verification. Replacing the reference sets changes the code space and requires re-encoding; fixed-reference reuse, as evaluated in Appendix F.6, avoids that change.

3 COLLISIONS, BUCKET LOADS, AND LOCALITY

The central quantity is the probability that two embeddings select the same reference in one round. Unless stated otherwise, probabilities condition on a fixed reference collection D. Rounds are iid, reference samples and their order are shared by all inputs, and parameters are fixed before drawing the rounds. For a fixed pair, define

$$p_\psi(x, y) = \mathbb{P}[h_t(x) = h_t(y)], \qquad F_{T,b}(p) = 1 - (1 - p^b)^{T/b}, \tag{4}$$

where the band rule requires $b \mid T$.

Proposition 3.1 (Finite-reference collisions with shared-order ties)**.** *Let D contain n indexed reference items, sampled uniformly without replacement in a uniform order. For a possible common winner z_j, let a_j count other items strictly closer at either endpoint. Of the remaining other items, let e_j count those tied with z_j at either endpoint, and let $v_j = n - 1 - a_j - e_j$ count those strictly farther at both. Under smallest-local-index tie breaking,*

$$p_\psi(x, y) = \frac{1}{\binom{n}{\psi}} \sum_{j=1}^{n} \frac{\binom{n-a_j}{\psi} - \binom{v_j}{\psi}}{e_j + 1}. \tag{5}$$

Impossible binomial coefficients are zero; repeated vectors remain separate indexed items.

A common winner must precede its sampled ties and avoid every strict competitor; Appendix A.1 gives the proof. Without ties, this reduces to the WTA law of Yagnik et al. (2011, Eq. 6) applied to negative distances.

From one round to candidate acceptance. Coordinate agreements are iid Bernoulli variables with parameter p_ψ. The usual disjoint-band amplification and binomial count tail give

$$\mathbb{P}[C(x, y)] = F_{T,b}(p_\psi(x, y)), \tag{6}$$

$$\mathbb{P}[A_m(x, y)] = G_{T,m}(p_\psi(x, y)), \quad G_{T,m}(p) = \sum_{j=m}^{T} \binom{T}{j} p^j (1 - p)^{T-j}. \tag{7}$$

Here C is a match in any complete band, and endpoints $p = 0, 1$ are interpreted by continuity. No extra hash transformation is used. Both functions are monotone and T-Lipschitz. Local neighborhoods can certify p_ψ within width δ, giving candidate-probability width at most $T\delta$ (Proposition A.1 in Appendix A.9). The empirical agreement $\widehat{p}_T = S_T/T$ is unbiased; Appendix A.3 gives its concentration bound. These are fixed-pair laws. Population acceptance averages $F_{T,b}(p_\psi)$ or $G_{T,m}(p_\psi)$ over pairs, rather than applying either function to the average agreement. Conditioning on parameters selected using the same random index need not preserve these identities; Appendix A.11.3 states the distinction and bounds that remain valid under such selection.

For N fixed database items, the number C_N of distinct directed non-self candidates satisfies

$$\mathbb{E}[C_N] = \sum_{i \neq j} \Phi(p_\psi(x_i, x_j)), \qquad \Phi \in \{F_{T,b}, G_{T,m}\}. \tag{8}$$

This is linearity of expectation, without independence between pairs. Uniform labels do not imply balanced fitted buckets. The following condition controls the latter directly.

Coverage and reference concentration control bucket loads. Fix a database $\mathcal{X}$ of N indexed items. Let $r_k(x)$ be its kth-smallest reference distance, for $1 \le k \le n$, and define

$$\mathcal{N}_k(x) = \{z \in D : d(x,z) \le r_k(x)\}, \quad R_k = \max_{z \in D} |\{x \in \mathcal{X} : z \in \mathcal{N}_k(x)\}|. \tag{9}$$

Every cutoff tie is included. The reverse-neighbor population R_k measures how many inputs can share a locally competitive reference, relating the load condition to hubness (Radovanović et al., 2010).

Proposition 3.2 (Maximum load from reference neighborhoods)**.** *For T uniform samples of ψ distinct reference items, let M_* be the largest coordinate bucket across all rounds. With $q_k = \binom{n-k}{\psi}/\binom{n}{\psi}$, for every integer $s \ge 1$,*

$$\mathbb{P}[M_* \ge R_k + s] \le \min\{1, TNq_k/s\}. \tag{10}$$

Each bucket contains at most R_k covered inputs plus the missed neighborhoods, whose expectation is at most Nq_k. Appendix B.4 proves the bound and gives data-only rank selection.

Continuous data can still produce extreme loads. Data dependence alone cannot make this condition hold. Proposition B.2 constructs distinct unit vectors from continuous distributions in a fixed dimension whose shared reference attracts most inputs. For example, with $N = 65{,}536$ in $\mathbb{R}^{257}$, $\psi = 1024$ and $T = 100$, its numerical bound gives a coordinate bucket containing at least 95.3% of the database with probability greater than 0.999 over the joint data and encoder draws. All coordinate buckets are nonempty almost surely. Thus ties, repeated vectors and empty buckets are unnecessary for severe concentration. Appendix B.5 gives the construction and proof; a separate growing-dimensional family establishes quadratic candidate work despite sparse true output.

When do near pairs collide and far pairs separate? For one reference sample Z, let $\Delta_Z(x) = d_{(2)}(x,Z) - d_{(1)}(x,Z)$ be its nearest-reference margin, and let $\nu_R(x) = \mathbb{P}[Z \cap B(x, R/2) = \varnothing]$, where B is a closed metric ball.

Theorem 3.3 (Near capture and far rejection)**.** *Let d be a metric, $\psi \ge 2$, and $0 < r \le R$. Restrict pairs to the target query and database domains. Suppose every such pair with $d(x,y) \le r$ has an endpoint u satisfying $\mathbb{P}[\Delta_Z(u) \le 2r] \le \eta$, and each endpoint of a pair with $d(x,y) > R$ satisfies $\nu_R(u) \le \nu$. Set $a = 1 - \eta$ and $c = \min\{1, 2\nu\}$, where $\eta, \nu \in [0,1]$. Choose either $(\mathcal{E}, \Phi) = (C, F_{T,b})$ or $(\mathcal{E}, \Phi) = (A_m, G_{T,m})$. Then*

$$d(x,y) \le r \implies \mathbb{P}[\mathcal{E}(x,y)] \ge \Phi(a) =: \alpha, d(x,y) > R \implies \mathbb{P}[\mathcal{E}(x,y)] \le \Phi(c) =: \beta. \tag{11}$$

The triangle inequality makes the nearest index stable whenever its margin exceeds $2r$. Conversely, a pair farther than R cannot share its nearest reference if both endpoints have a reference within $R/2$. These give $p_\psi \ge a$ and $p_\psi \le c$, respectively; both $F_{T,b}$ and $G_{T,m}$ are nondecreasing, proving the theorem.

Increasing ψ improves reference coverage but can reduce the margin, so locality need not improve monotonically. When $a > c$, banding trades far rejection against near capture with $L = T/b$ (Appendix A.5); if $c < m/T < a$, counting satisfies the exponential error bounds of Appendix A.11.1. These guarantees require the stated distribution-dependent separation. For unit vectors, $\|x - y\|_2^2 = 2(1 - \langle x, y\rangle)$ translates the metric thresholds to cosine similarity.

What does candidate acceptance imply? The reverse implication also depends on the distribution of pairs. Let (X, Y) be drawn from these domains independently of encoder randomness, with $\pi_N = \mathbb{P}[d(X,Y) \le r]$ and $\pi_F = \mathbb{P}[d(X,Y) > R]$.

Corollary 3.4 (Precision of collision candidates)**.** *Under Theorem 3.3, if $\pi_N \alpha > 0$, then*

$$\mathbb{P}[d(X,Y) \le R \mid \mathcal{E}(X,Y)] \ge \frac{\pi_N \alpha}{\pi_N \alpha + \pi_F \beta}. \tag{12}$$

This bound permits arbitrary pair mass in $(r, R]$. High candidate precision requires far acceptance to be rare relative to near-pair prevalence. The statement averages over the specified pair draw and encoder randomness; it does not certify every realized bucket or equate geometric proximity with semantic duplication. Full proofs appear in Appendices A and A.11. Appendix A.10 gives an exact uniform-circle example with a strictly decreasing collision curve and explicit candidate precision under a uniform pair prior.

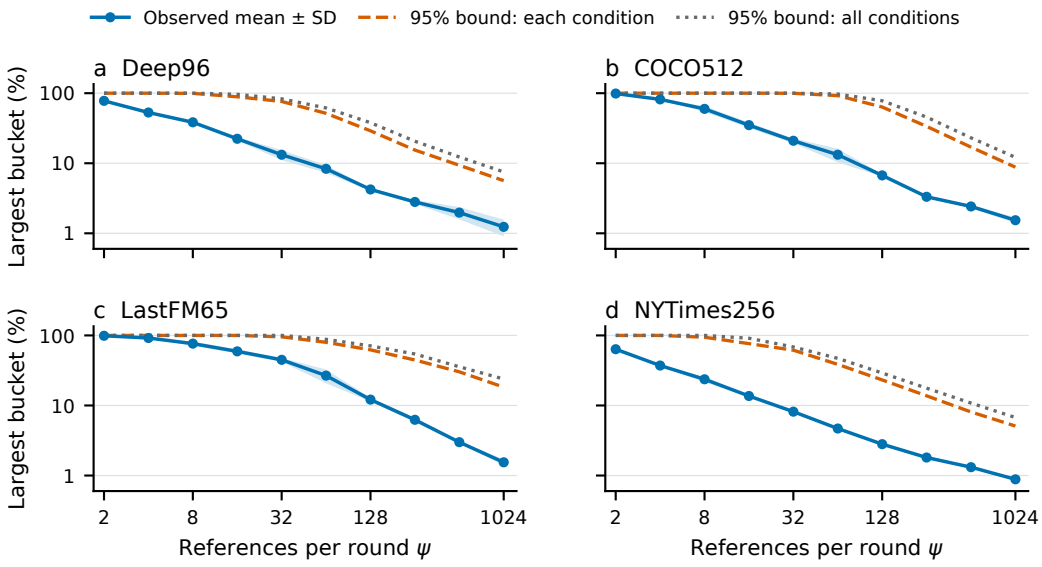

Figure 2: **Larger reference sets reduce measured concentration; the load bounds remain conservative.** Largest coordinate bucket as a fraction of each 100k database at $T = 100$. Solid curves show means across three fits, with sample-standard-deviation shading. Upper bounds from Equation (83) use failure probability 0.05 per dataset/resolution/fit, or 0.05 jointly over all 120 conditions. The bounds condition on the fixed score matrix; they are not confidence intervals for the plotted means.

From locality to complete workload. Let M_R count directed database pairs within distance R. If every farther pair has candidate probability at most β, then $\mathbb{E}[C_N] \le M_R + \{N(N-1) - M_R\}\beta$ (Appendix B). Hence $M_R = O(N)$ and $\beta = O(1/N)$ suffice for linear expected candidate volume; a fixed signature length alone does not. This connects locality to the task workload: graph construction and joins expose all database pairs, whereas retrieval exposes only the requested query–database pairs. Signature reuse saves repeated encoding, but it does not remove verification work or the cost of producing a dense output.

4 Experiments

We evaluate pairwise calibration, fitted loads, and complete similarity tasks. Validation selects all operating points[1]. New-query confirmation is distinguished from controls on previously observed queries, and candidate retention from returned recall. Appendix C gives protocols. Table subscripts report ± sample SD across three fits; superscripts retain their stated quality flags. Bold identifies the best eligible displayed means, including ties, under each table's comparison rule; it does not imply statistical significance.

4.1 When does data dependence control bucket concentration?

We evaluate Proposition 3.2 on Deep96, COCO512, LastFM65 and NYTimes256, each with 100,000 database items and a 16,384-item data reference pool. We use $T = 100$, all ten $\psi \in \{2, 4, \ldots, 1024\}$, and three independent fits. Each fit's statistic is the largest coordinate bucket over all rounds. The reference-neighborhood profile chooses a rank before observing these fitted loads. The bound and hash evaluation share fixed FP64 scores; this diagnostic does not replace the FP32 task comparisons.

At $\psi = 2$, the worst observed COCO and LastFM buckets each contain more than 99% of their database, despite data-derived references. At $\psi = 1024$, mean maximum loads range from 0.883% to 1.544%, whereas the primary upper bounds range from 5.088% to 18.275%. The primary certificate

[1]Source code is available in `https://anonymous.4open.science/r/SMinHash/`.

Table 1: **Data references reduce candidates; native vector indexes query faster.** Means over three fits on 100k items and 1,000 new queries per dataset. Data denotes SMinHash with data references; Sphere uses spherical references. Query times cover all 1,000 queries. Superscripts count fits below 95% returned Recall@10 (omitted when zero). Bold recall values are the larger reference-family means. Bold candidate minima require both reference sources to meet the target in every fit; bold query minima require that family to do so.

	Recall@10 (%)		Candidates (%)		Query (ms)			
Dataset	Data	Sphere	Data	Sphere	Data	Sphere	Flat	IVF-Flat
Deep96	96.607	**97.500**	**3.146**	12.094	72.218	62.779	13.673	**7.758**
COCO512	95.767[1]	**97.523**	0.922	30.908	83.595[1]	209.738	21.496	**5.887**
GloVe25	95.890	**96.970**	**3.001**	6.006	91.399	94.098	**13.768**	15.616
GloVe100	**98.223**	97.610	**17.056**	31.323	66.008	78.331	**7.368**	10.255
LastFM65	**99.743**	98.883	**1.417**	94.221	73.275	293.548	**7.032**	8.729
NYTimes256	97.327	**98.570**	**26.467**	52.458	107.656	179.948	**11.462**	30.323
DBpedia1536	**98.893**	96.360	**3.995**	32.694	114.371	671.209	**33.870**	48.120[3]
COYO2048	95.027[2]	**97.417**	2.522	28.582	87.156[2]	606.529	**45.419**	39.141[3]

improves on the trivial 100% bound in 28 of 40 dataset/resolution settings; the simultaneous certificate does so in 25. The bound is informative in some settings but remains conservative; its posting-work consequence does not certify the task budget (Appendix B.4).

The next comparison tests whether reference adaptation reduces useful candidate work after independent tuning. Appendix B.4 gives the score and confidence conventions; Appendices E.6 and E.9 examine collision calibration and local probability intervals.

4.2 Does reference adaptation reduce complete query cost?

We compare data-derived references with independent spherical references within the same nearest-index encoder. Both receive the same validation budget: $T = 100$, ten $\psi \in \{2, 4, \ldots, 1024\}$ values and all nine complete band widths dividing 100. Each family minimizes candidates subject to retaining 95% of validation top-10 neighbors. Choices are fixed before evaluating 1,000 new query IDs on each of eight 100k databases. Table 1 summarizes three fits; the full returned-recall and dispersion results are in Table 13.

Data references pass fewer candidates in all 21 paired fits where both sources attain 95% returned Recall@10, out of 24 pairs. They also have lower mean candidate fractions on all eight datasets and lower mean query time on seven. LastFM's fractions are 1.417% versus 94.221%, with both sources reaching the target in every fit. The effect persists after independent tuning, but does not ensure test recovery: data references miss the target in one COCO fit and two COYO fits.

Complete query cost includes encoding, lookup, union, scoring and output assembly on resident inputs, in requests of 32. Both reference families score all candidates in FP64; native GPU Flat and IVF-Flat use FP32 and their native tie rules. Host copies are excluded for all four. Data references build faster than IVF-Flat in all dataset means, but Flat queries faster on all eight; Deep is also a case where fewer candidates do not beat spherical references. IVF-Flat misses the quality target on DBpedia and COYO in every fit. Thus adaptation reduces useful candidate work without guaranteeing the fastest full pipeline. Appendix E.1 distinguishes this comparison from the fixed-parameter source intervention and gives measurement scopes.

4.3 Can SMinHash screen semantic matches?

We use MRPC, PAWS-Wiki, and QQP pairs (Dolan & Brockett, 2005; Zhang et al., 2019; Wang et al., 2019; nyu-mll, 2024), with frozen normalized MiniLM embeddings shared by all methods. Development data select the cosine threshold and one configuration per family, requiring at least 99% retention of cosine-positive development pairs in each of three hashing seeds. Candidate fractions use a separate fixed cross-product probe, measuring background screening cost independently of the labeled pairs in Table 2.

Table 2: **Selective gates preserve most cosine-positive semantic pairs.** Mean$_{\pm SD}$ over three fits; exact cosine is deterministic. SMinHash uses $T = 100$ and direct banding; Hyperplane matches its nominal bit capacity. Retain measures cosine-positive pair recovery, and Probe cand. measures a separate cross-product probe, not a full-corpus join. Bold marks the highest F1 and retention; candidate minima require mean retention at least 99%. Displayed ties are included. F1 is approximately preserved; small differences are not treated as classifier improvements.

Dataset	Method	F1 (%)	Probe cand. (%)	Retain (%)
MRPC	Exact cosine	82.033	100.000	**100.000**
	Hyperplane (bits)	$82.024_{\pm 0.033}$	$13.167_{\pm 0.171}$	$99.907_{\pm 0.107}$
	SMinHash	$\mathbf{82.075}_{\pm 0.055}$	$\mathbf{3.311}_{\pm 0.014}$	$99.510_{\pm 0.070}$
PAWS-Wiki	Exact cosine	61.369	100.000	**100.000**
	Hyperplane (bits)	$\mathbf{61.392}_{\pm 0.021}$	$\mathbf{0.340}_{\pm 0.003}$	$99.428_{\pm 0.026}$
	SMinHash	$61.254_{\pm 0.078}$	$0.230_{\pm 0.012}$	$98.989_{\pm 0.095}$
QQP	Exact cosine	74.456	100.000	**100.000**
	Hyperplane (bits)	$74.456_{\pm 0.000}$	$12.601_{\pm 0.224}$	$\mathbf{100.000}_{\pm 0.000}$
	SMinHash	$\mathbf{74.508}_{\pm 0.019}$	$\mathbf{0.338}_{\pm 0.008}$	$99.482_{\pm 0.023}$

On QQP, SMinHash retains 99.48% of cosine-positive pairs while passing 0.338% of probe pairs to the cosine verifier, preserving the exact cosine classifier's F1. Capacity-matched hyperplane signatures pass 12.601% of probe pairs while retaining 100% of cosine positives. PAWS retention falls below the development target, which is not a held-out guarantee. Screening also inherits the encoder's limitations: exact cosine barely exceeds the all-positive classifier on PAWS, and the encoder's Quora training data limit QQP's evidence for unseen-data generalization. These are semantic-rule screening experiments; Appendix C details their scope and data overlap.

Figure 3 compares both retention targets on all three datasets, including similar candidate fractions. Product blocking on 500 official WDC matches in a billion-pair space uses fewer candidates than Hyperplane LSH; top-R search remains more selective (Appendix D.3).

4.4 Agreement thresholds and representation geometry

An additional new-query comparison uses the same signatures with complete collision-count gates against Hyperplane1000 (Appendix D.1). On LastFM, SMinHash retains 95.82% of true neighbors with 0.332% candidates, versus 95.26% and 45.298% for the control. A centering intervention reduces the control to 0.732% candidates but 94.830% retention, linking much of the gap to the asymmetric user–item representation (Appendix E.3). Across the eight datasets, each family misses its 95% validation target in 12 of 24 new-query fits. Complete retrieval on these fixed queries retains the same limitation: Flat and IVF-Flat remain faster on LastFM, and SMinHash's faster times than Flat on DBpedia and COYO occur below the quality target (Appendix F.2).

4.5 Graphs and threshold joins

The same signature supports all-node queries and complete threshold joins. Parameters are selected using 1,024 validation vertices for the corresponding task; thresholds use a fixed background-density calibration. Table 3 compares complete joins with GPU-assisted IVF-Flat range search and exhaustive verification on four 100k databases. IVF validates probe counts and training budgets for construction plus one join; SMinHash selects execution batches with its hash parameters and candidate rule unchanged. All pipelines run serially on the same GPU. Appendix C.3.5 gives the range protocol; Appendix D.8 reports graph construction against working ANN baselines.

SMinHash completes all four joins with 95.8–98.3% edge recall and 0.14–0.33% candidates. On COCO, its lower construction cost gives 0.734 seconds in total versus 1.046 for GPU-assisted IVF and 13.697 for exhaustive verification. With the expanded IVF budget search, SIFT totals are similar (0.566 versus 0.567 seconds), and IVF has the lower total on NYTimes (0.750 versus 0.984). IVF's mean query phase is faster on all four datasets. One Deep IVF fit misses the 0.95 recall target; every

Table 3: **Construction cost affects the complete-join ranking.** Mean post-selection construction-plus-join cost over three fits on 100k vectors; tuning is excluded. Exhaustive recall is one. †IVF cost includes GPU training/search, uncapped CPU overflow and FP64 verification. Bold recall marks the higher approximate-method mean; bold total marks the minimum among methods reaching 0.95 recall in every fit. Superscripts count failures. Achieved recalls differ.

	SMinHash		IVF-GPU†		Exhaustive
Dataset	Recall	Total (s)	Recall	Total (s)	Total (s)
COCO512	0.958^{0}	$\mathbf{0.734}$	$\mathbf{0.969}^{0}$	1.046	13.697
Deep96	0.961^{0}	$\mathbf{0.383}$	$\mathbf{0.964}^{1}$	0.434	3.690
SIFT128	0.983^{0}	$\mathbf{0.566}$	$\mathbf{0.992}^{0}$	0.567	5.008
NYTimes256	$\mathbf{0.964}^{0}$	0.984	0.954^{0}	$\mathbf{0.750}$	7.520

SMinHash fit meets it. Thus low verification volume supports efficient joins, but the cost advantage depends on construction budget, workload and achieved recall.

For graphs, SMinHash completes COCO, Deep and SIFT with 95.1–96.7% audited neighbor recall and 1.10–3.12% candidate pairs, but exceeds the posting budget on GIST and NYTimes. Flat and IVF-Flat have lower graph costs wherever SMinHash completes. Geometric LSH exceeds its posting budget in all four joins; the appendix retains its graph failures and infeasible selections.

With frozen 100k selections, references and query IDs, Deep grows to 9.99 million items: SMinHash recovers 99.18% of true top-ten neighbors while verifying 3.096% of the database per query. Complete costs appear in Appendix F.5; no scale-specific retuning is used.

5 Related Work

Minimum selection and reference agreement. WTA already relates extreme-index selection to set MinHash and discusses distance-derived features (Yagnik et al., 2011, Secs. 4.3–4.4). With matched reference samples, a full-data reference pool and one row per band, SMinHash has the same candidate rule as single-probe DFLSH, up to the choice at distance ties (Kang & Jung, 2012, Sec. 2). Its locality analysis uses a local-uniformity model and its load comparison is empirical (Kang & Jung, 2012, Sec. 3). Our fixed-pool analysis separates coverage from reverse-neighbor concentration, using established hubness statistics (Radovanović et al., 2010). Collision counting and multi-index prefiltering are inherited tools (Gan et al., 2012; Norouzi et al., 2012).

Embedding neighborhood tasks. HNSW, product quantization and GPU vector search supply strong retrieval comparators (Malkov & Yashunin, 2020; Jégou et al., 2011; Johnson et al., 2021; Douze et al., 2024). NN-Descent targets graphs (Dong et al., 2011), SemDeDup uses restricted embedding comparisons for curation (Abbas et al., 2023), and SC-Block learns representations for entity blocking (Brinkmann et al., 2024). Representation analyses connect anisotropy and bucket entropy to retrieval (Rodríguez-Betancourt & Casasola-Murillo, 2026). Our reference-source intervention holds the embeddings fixed; complete-cost comparisons distinguish selectivity from execution cost. Geometric matches remain distinct from semantic equivalence.

6 Conclusion

Semantic MinHash brings the minimum-selection and bucket-based processing of MinHash and LSH to semantic embeddings through constructed reference sets. Each embedding is represented by its distances to shared sampled points, and the minimum-distance indices form its reusable hash signature. Our finite-reference collision law, conditional locality and candidate-precision guarantees, and maximum-load analysis provide a theoretical basis for relating these signatures to embedding similarity and candidate work. Experiments demonstrate that data-derived references improve selectivity over a spherical control, preserve most cosine-positive pairs in semantic screening, and support graph and join processing with reduced verification volume.

AI USE STATEMENT

Generative AI assisted with language editing, literature and reference checks, and the checking of theoretical statements and proofs. This introduced no new experimental measurements. The authors retain responsibility for the final manuscript, including verification of all AI-assisted content.

ETHICS STATEMENT

This work concerns MinHash. No new human-subject data were collected. MiniHash can amplify representation bias, so deployment in consequential settings requires dataset-specific fairness, privacy, and error analysis.

REFERENCES

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023. URL https://arxiv.org/abs/2303.09540.

Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. ANN-Benchmarks: A benchmarking tool for approximate nearest neighbor algorithms. *Information Systems*, 87:101374, 2020. doi: 10.1016/j.is.2019.02.006. URL https://doi.org/10.1016/j.is.2019.02.006.

Alexander Brinkmann, Roee Shraga, and Christian Bizer. WDC Block: A blocking benchmark. Web Data Commons, University of Mannheim, 2023. URL https://www.webdatacommons.org/largescaleproductcorpus/wdc-block/. Medium data size, large development set.

Alexander Brinkmann, Roee Shraga, and Christina Bizer. SC-Block: Supervised contrastive blocking within entity resolution pipelines. In *The Semantic Web*, volume 14664 of *Lecture Notes in Computer Science*, pp. 121–142. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-60626-7_7. URL https://link.springer.com/chapter/10.1007/978-3-031-60626-7_7. ESWC 2024.

Andrei Z. Broder. On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of Sequences*, pp. 21–29. IEEE, 1997. URL https://www.cs.princeton.edu/courses/archive/spring13/cos598C/broder97resemblance.pdf.

Moses S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pp. 380–388. ACM, 2002. URL https://www.cs.princeton.edu/courses/archive/spring04/cos598B/bib/CharikarEstim.pdf.

Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the Twentieth Annual Symposium on Computational Geometry*, pp. 253–262. ACM, 2004. doi: 10.1145/997817.997857. URL https://doi.org/10.1145/997817.997857.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*, 2005. URL https://aclanthology.org/I05-5002/.

Wei Dong, Moses Charikar, and Kai Li. Efficient K-nearest neighbor graph construction for generic similarity measures. In *Proceedings of the 20th International Conference on World Wide Web*, 2011. URL https://www.cs.princeton.edu/cass/papers/www11.pdf.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The Faiss library, 2024. URL https://arxiv.org/abs/2401.08281.

Faiss contributors. Faiss FAQ, 2025. URL https://github.com/facebookresearch/faiss/wiki/FAQ. Accessed September 24, 2026.

Junhao Gan, Jianlin Feng, Qiong Fang, and Wilfred Ng. Locality-sensitive hashing scheme based on dynamic collision counting. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pp. 541–552, 2012. doi: 10.1145/2213836.2213898. URL https://cse.hkust.edu.hk/~wilfred/paper/sigmod12.pdf.

Nicholas J. Higham and Theo Mary. A new approach to probabilistic rounding error analysis. *SIAM Journal on Scientific Computing*, 41(5):A2815–A2835, 2019. doi: 10.1137/18M1226312.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Sergey Ioffe. Improved consistent sampling, weighted Minhash and L1 sketching. In *IEEE International Conference on Data Mining*, 2010. URL https://research.google/pubs/improved-consistent-sampling-weighted-minhash-and-l1-sketching/.

H. Jégou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011. doi: 10.1109/TPAMI.2010.57. URL https://doi.org/10.1109/TPAMI.2010.57.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2021. doi: 10.1109/TBDATA.2019.2921572. URL https://doi.org/10.1109/TBDATA.2019.2921572.

Byungkon Kang and Kyomin Jung. Robust and efficient locality sensitive hashing for nearest neighbor search in large data sets. In *NIPS Workshop on Big Learning*, 2012. URL https://milab.snu.ac.kr/pub/2012BigLearn.pdf.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking, 2026. URL https://arxiv.org/abs/2601.04720v2. Version 2, January 19, 2026.

Qin Lv, William Josephson, Zhe Wang, Moses Charikar, and Kai Li. Multi-probe LSH: Efficient indexing for high-dimensional similarity search. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, pp. 950–961, 2007. URL https://www.vldb.org/conf/2007/papers/research/p950-lv.pdf.

Yury 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. URL https://doi.org/10.1109/TPAMI.2018.2889473.

Behnam Neyshabur and Nathan Srebro. On symmetric and asymmetric LSHs for inner product search. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 1926–1934. PMLR, 2015. URL https://proceedings.mlr.press/v37/neyshabur15.html.

Mohammad Norouzi, Ali Punjani, and David J. Fleet. Fast search in Hamming space with multi-index hashing. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. URL https://www.cs.toronto.edu/~fleet/research/Papers/chunkingCVPR12.pdf.

nyu-mll. GLUE QQP public data files. Hugging Face dataset repository, 2024. URL https://huggingface.co/datasets/nyu-mll/glue/tree/bcdcba79d07bc864c1c254ccfcedcce55bcc9a8c/qqp. Parquet distribution committed 4 January 2024; pinned revision bcdcba79d07bc864c1c254ccfcedcce55bcc9a8c; accessed 17 September 2026.

OpenAI. New embedding models and API updates. OpenAI product announcement, January 25, 2024, January 2024. URL https://openai.com/index/new-embedding-models-and-api-updates/. Model source for text-embedding-3-large.

Qdrant. DBpedia entities: OpenAI text-embedding-3-large, 1536 dimensions, 1m. Hugging Face dataset, February 2024. URL https://huggingface.co/datasets/Qdrant/dbpedia-entities-openai3-text-embedding-3-large-1536-1M/tree/4a9731217921bc476a0f03544f11f22ae4903fa5. Frozen dataset revision 4a9731217921bc476a0f03544f11f22ae4903fa5; precomputed vectors published by Qdrant.

Qdrant. Coyo-Vector-Embeddings (Coyo-VE). Hugging Face dataset, 2026. URL https://huggingface.co/datasets/Qdrant/Coyo-VE/tree/acb2de723c84485b1992edea27ffe992fa9ed20f. Frozen dataset revision acb2de723c84485b1992edea27ffe992fa9ed20f; joint image-text vectors published by Qdrant.

Xiaoyu Qin, Kai Ming Ting, Ye Zhu, and Vincent C. S. Lee. Nearest-neighbour-induced isolation similarity and its impact on density-based clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4755–4762, 2019. doi: 10.1609/aaai.v33i01.33014755. URL https://ojs.aaai.org/index.php/AAAI/article/view/4402.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(86):2487–2531, 2010. URL https://www.jmlr.org/papers/v11/radovanovic10a.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Esteban Rodríguez-Betancourt and Edgar Casasola-Murillo. Geometric analysis of self-supervised vision representations for semantic image retrieval, 2026. URL https://arxiv.org/abs/2604.24469.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1804.07461.

Yang Xu and Kai Ming Ting. Voronoi diagram encoded hashing. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2025. URL https://ecmlpkdd-storage.s3.eu-central-1.amazonaws.com/preprints/2025/research/preprint_ecml_pkdd_2025_research_223.pdf. Conference-hosted preprint; also arXiv:2508.02266.

Jay Yagnik, Dennis Strelow, David A. Ross, and Ruei-sung Lin. The power of comparative reasoning. In *2011 International Conference on Computer Vision*, pp. 2431–2438, 2011. doi: 10.1109/iccv.2011.6126527. URL https://research.google/pubs/the-power-of-comparative-reasoning/.

Xiao Yan, Jinfeng Li, Xinyan Dai, Hongzhi Chen, and James Cheng. Norm-ranging LSH for maximum inner product search. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/b60c5ab647a27045b462934977ccad9a-Abstract.html.

Yuan Zhang, Jason Baldridge, and Luheng He. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1298–1308. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1131. URL https://aclanthology.org/N19-1131/.

Zhibo Zhang, Yang Xu, Kai Ming Ting, and Cam-Tu Nguyen. LLMs meet isolation kernel: Lightweight, learning-free binary embeddings for fast retrieval. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13601–13623. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.666. URL https://aclanthology.org/2026.findings-acl.666/.

A COLLISION PROBABILITIES AND LOCALITY

Each round selects a uniformly ordered subset of ψ reference items without replacement from a fixed collection $D = (z_1, \ldots, z_n)$. Different rounds are independent. Inputs share the reference samples, their order, and the smallest-local-index tie rule. Unless an iid population model is explicitly stated, all probabilities condition on D. The signature consists of the $T = Lb$ nearest indices themselves.

A.1 EXACT FINITE-REFERENCE PROBABILITY, INCLUDING DISTANCE TIES

The reference-feature correspondence. For ordered sampled IDs $I = (i_0, \ldots, i_{\psi-1})$, let $s_D(x)_i = -d(x, z_i)$. With first-position tie breaking,

$$\operatorname*{first\,argmax}_{0 \le j < \psi} s_D(x)_{i_j} = \operatorname*{first\,argmin}_{0 \le j < \psi} d(x, z_{i_j}) = h_I(x).$$

Thus applying WTA to these reference features gives the same index for every ordered sample, even with ties. A uniform permutation's first ψ IDs form precisely a uniform ordered sample without replacement. This correspondence interprets the existing encoder; it adds no hash. For strict distance orders, the lower-score intersection in Yagnik et al. (2011, Eq. 6) is

$$\{i : s_D(x)_i < s_D(x)_j,\ s_D(y)_i < s_D(y)_j\} = \{i : d(x, z_i) > d(x, z_j),\ d(y, z_i) > d(y, z_j)\},$$

whose size is $n - 1 - a_j$. Substitution gives the existing strict-order collision formula

$$p_\psi(x, y) = \frac{1}{\binom{n}{\psi}} \sum_{j=1}^{n} \binom{n - 1 - a_j}{\psi - 1}. \tag{13}$$

The proof of Proposition 3.1 below integrates the shared sampled-order tie convention instead of imposing strict orders or holding a global tie priority fixed.

Fix x, y. For a possible common winner z_j, partition the other reference items into three groups:

$$A_j = \{i \ne j : d(x, z_i) < d(x, z_j) \text{ or } d(y, z_i) < d(y, z_j)\}, \tag{14}$$
$$E_j = \{i \notin A_j \cup \{j\} : d(x, z_i) = d(x, z_j) \text{ or } d(y, z_i) = d(y, z_j)\}, \tag{15}$$
$$V_j = [n] \setminus (A_j \cup E_j \cup \{j\}). \tag{16}$$

Write $a_j = |A_j|$, $e_j = |E_j|$, and $v_j = |V_j| = n - 1 - a_j - e_j$. An item in A_j defeats j at at least one endpoint regardless of sample order. An item in E_j defeats j at an endpoint only if it occurs before j. Items in V_j are strictly farther from both endpoints. Thus, if a sample contains j, excludes A_j, and includes k members of E_j, it has j as common winner exactly when j comes first among these $k + 1$ tied contenders. Uniform ordering gives probability $1/(k + 1)$. Counting the possible unordered subsets yields

$$p_\psi(x, y) = \frac{1}{\binom{n}{\psi}} \sum_{j=1}^{n} \sum_{k=0}^{\psi-1} \frac{\binom{e_j}{k}\binom{v_j}{\psi-1-k}}{k + 1}. \tag{17}$$

The common-winner events are disjoint across j, including ties, because the local-order rule chooses one index at each endpoint. Impossible binomial coefficients are zero throughout.

Using $\binom{e}{k}/(k + 1) = \binom{e+1}{k+1}/(e + 1)$ and Vandermonde's identity gives an equivalent compact formula:

$$\sum_{k=0}^{\psi-1} \frac{\binom{e_j}{k}\binom{v_j}{\psi-1-k}}{k + 1} = \frac{1}{e_j + 1} \sum_{\ell=1}^{\psi} \binom{e_j + 1}{\ell}\binom{v_j}{\psi - \ell} = \frac{\binom{n-a_j}{\psi} - \binom{v_j}{\psi}}{e_j + 1}. \tag{18}$$

Substitution into Equation (17) proves Equation (5). This expression integrates the sampled-order tie rule; it does not hold a global tie priority fixed across rounds. With no ties, $e_j = 0$ and Pascal's identity reduces the contribution to $\binom{n-1-a_j}{\psi-1}$, proving Equation (13). Duplicate reference vectors are still separate indexed items, so the formula also covers repeated coordinates in D. At $\psi = 1$, every pair selects the only sampled item, giving $p_1(x, y) = 1$.

An iid population alternative. If the ψ references within a round are iid from a distribution μ instead, define the strict competitor region

$$U_{x,y}(z) = \{w : d(x,w) < d(x,z) \text{ or } d(y,w) < d(y,z)\}. \tag{19}$$

When endpoint distances have no ties almost surely, independence and exchangeability give

$$p_\psi(x,y) = \psi \int [1 - \mu(U_{x,y}(z))]^{\psi-1}\, d\mu(z). \tag{20}$$

Indeed, after conditioning on one reference being z, every other reference must avoid $U_{x,y}(z)$ for that index to win at both inputs. For discrete distributions or other tied cases, define

$$a(z) = \mathbb{P}_W[d(x,W) > d(x,z),\ d(y,W) > d(y,z)], \tag{21}$$
$$c(z) = \mathbb{P}_W[d(x,W) \geq d(x,z),\ d(y,W) \geq d(y,z)]. \tag{22}$$

For index j to be the common minimum under smallest-index tie breaking, all j earlier references must be strictly farther at both endpoints, whereas later references may tie. Consequently,

$$p_\psi(x,y) = \sum_{j=0}^{\psi-1} \int a(z)^j c(z)^{\psi-1-j}\, d\mu(z). \tag{23}$$

Zero exponents contribute one, including when the base is zero. The iid formula and finite-pool formula analyze different sampling designs; our implementation uses the latter.

A.2 The classical set-MinHash special case

The WTA binary specialization recovers ordinary MinHash with a complete permutation (Yagnik et al., 2011, Sec. 4.4). The following distance representation makes that known connection explicit for our encoder and identifies the effect of sampling fewer references. It uses a specified dictionary, not an assumption about arbitrary data.

Let A, B be nonempty subsets of $[n]$, $n \geq 2$, with indicator vectors $x_A, x_B \in \{0,1\}^n$. Choose $D = (e_1, \ldots, e_n)$ and Euclidean distance. Squaring preserves the nearest-reference order, and

$$\|x_A - e_j\|_2^2 = |A| + 1 - 2\mathbf{1}\{j \in A\}. \tag{24}$$

For an ordered sample $I = (i_0, \ldots, i_{\psi-1})$, the first-minimum rule therefore gives

$$h_I(x_A) = \begin{cases} \min\{k : i_k \in A\}, & A \cap \{i_0, \ldots, i_{\psi-1}\} \neq \varnothing, \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

At $\psi = n$, absence is impossible. The output is the permuted rank of the first member of A, exactly ordinary MinHash. Two such ranks agree if and only if the first element of $A \cup B$ lies in $A \cap B$. Uniform ordering then gives $p_n(x_A, x_B) = |A \cap B|/|A \cup B| = J(A,B)$. The local rank and winning reference ID are equivalent collision keys for a shared permutation.

For partial sampling, the fallback in Equation (25) changes this identity. Write $a = |A \setminus B|$, $b = |B \setminus A|$, $c = |A \cap B|$, $u = a + b + c$, $J = c/u$, and $q_0 = \binom{n-u}{\psi}/\binom{n}{\psi}$. Uniformly ordered sampling without replacement gives the exact law

$$p_\psi(x_A, x_B) = J(1 - q_0) + q_0 + \frac{a}{n}\frac{\binom{n-|B|-1}{\psi-1}}{\binom{n-1}{\psi-1}} + \frac{b}{n}\frac{\binom{n-|A|-1}{\psi-1}}{\binom{n-1}{\psi-1}}. \tag{26}$$

Here $1 \leq \psi \leq n$, including the diagnostic boundary $\psi = 1$; impossible binomial coefficients are zero, including negative upper arguments. To prove the formula, partition collisions into four disjoint events. If both sets are represented, their first sampled members agree exactly when the first sampled element of their union is in their intersection. Conditional on sampling the union, that first element is uniform within it, giving $J(1 - q_0)$. If neither set is represented, both hashes are zero, giving q_0. If B alone is absent, collision requires $i_0 \in A \setminus B$; this has probability a/n, and the remaining $\psi - 1$ items must avoid B among the other $n - 1$ references. This gives the third term. Interchanging A, B gives the fourth term, exhausting the possibilities.

In particular,

$$p_\psi - J = (1-J)q_0 + \frac{a\binom{n-|B|-1}{\psi-1} + b\binom{n-|A|-1}{\psi-1}}{n\binom{n-1}{\psi-1}} \geq 0.$$

For $n = 3$, $A = \{1\}$, $B = \{2\}$ and $\psi = 2$, this yields $p_2 = 1/3$ although $J = 0$. At $\psi = n$ the extra terms vanish; at $\psi = 1$ every hash is zero. These identities specialize the finite tie law to binary indicator inputs. They explain the classical connection and its sampling boundary, without transferring Jaccard calibration to general embeddings.

A.3 Candidate amplification and concentration

For a fixed pair, let $I_t = \mathbf{1}\{h_t(x) = h_t(y)\}$. Independent identically distributed reference rounds imply

$$I_t \overset{\text{iid}}{\sim} \text{Bernoulli}(p_\psi), \qquad M = \sum_{t=1}^{T} I_t \sim \text{Binomial}(T, p_\psi). \tag{27}$$

The event for band ℓ is $C_\ell = \bigcap_{t=(\ell-1)b+1}^{\ell b}\{I_t = 1\}$. Independence within a band and disjointness of the coordinate sets across bands give

$$\mathbb{P}[C_\ell] = p_\psi^b, \qquad \mathbb{P}[C^c] \quad = \mathbb{P}\left[\bigcap_{\ell=1}^{L} C_\ell^c\right] = \prod_{\ell=1}^{L}(1-p_\psi^b) = (1-p_\psi^b)^L. \tag{28}$$

Taking the complement proves Equation (6). This argument applies directly to the coordinates returned by Equation (1); there is no conditional Jaccard step. Monotonicity follows because both $p \mapsto p^b$ and $s \mapsto 1-(1-s)^L$ are nondecreasing on $[0,1]$.

Since $\widehat{p}_T = M/T$,

$$\mathbb{E}[\widehat{p}_T] = p_\psi, \qquad \text{Var}(\widehat{p}_T) = \frac{p_\psi(1-p_\psi)}{T}. \tag{29}$$

Hoeffding's inequality applied to the bounded independent variables I_t gives

$$\mathbb{P}[|\widehat{p}_T - p_\psi| \geq \epsilon] \leq 2e^{-2T\epsilon^2}. \tag{30}$$

For a fixed collection of m pairs, a union bound gives simultaneous error at most

$$\max_{i\in[m]} |\widehat{p}_T(x_i,y_i) - p_\psi(x_i,y_i)| \leq \sqrt{\frac{\log(2m/\delta)}{2T}} \tag{31}$$

with probability at least $1-\delta$. Pairs sharing an index need not be independent: the union bound requires only the fixed-pair result. The set of pairs must be fixed independently of the tested encoder, or the analysis must cover all pairs from which it is selected.

The complete-event formula also requires independent rounds and disjoint bands. Reusing the same reference sample in multiple coordinates would invalidate this factorization. If band lengths differ, $b_\ell \geq 1$ and $\sum_\ell b_\ell = T$, the formula becomes $1 - \prod_\ell (1-p_\psi^{b_\ell})$. Our experiments use equal complete bands with $b \mid T$.

A.4 Proof of Theorem 3.3

Near pairs: stability of the nearest index. Suppose $d(x,y) \leq r$ and $\Delta_Z(x) > 2r$. Let z_* be the unique nearest reference at x. For every other sampled reference z, the triangle inequality gives

$$d(y,z) - d(y,z_*) \geq d(x,z) - d(x,z_*) - 2d(x,y) \geq \Delta_Z(x) - 2r > 0. \tag{32}$$

Thus z_* also wins at y. The same argument works with the endpoints interchanged, so the stated endpoint assumption implies

$$p_\psi(x,y) \geq \mathbb{P}[\Delta_Z(u) > 2r] \geq 1-\eta = a. \tag{33}$$

The strict margin handles ties without a general-position assumption. The proof uses metric distances: squared Euclidean distance induces the same nearest index but must not replace the metric in the margin.

The sufficient stability probability can be counted exactly for a fixed reference pool. Let $w_j(g) = |\{i : d(x, z_i) > d(x, z_j) + g\}|$, where $g \geq 0$. A sample has z_j as minimum and margin greater than g exactly when it contains j and its other $\psi - 1$ items belong to this strictly farther set. Therefore, for $\psi \geq 2$,

$$\mathbb{P}[\Delta_Z(x) > g] = \frac{\sum_{j=1}^{n} \binom{w_j(g)}{\psi-1}}{\binom{n}{\psi}}. \tag{34}$$

This identity allows a direct diagnostic of the sufficient margin condition without treating an empirical lower confidence limit as an exact bound.

Far pairs: reference coverage. For $d(x, y) > R$, let $E_u = \{Z \cap B(u, R/2) = \varnothing\}$. If neither E_x nor E_y occurs, both nearest distances are at most $R/2$. A common winning reference would then imply

$$d(x, y) \leq d(x, z_*) + d(z_*, y) \leq R, \tag{35}$$

a contradiction. Consequently,

$$\{h_Z(x) = h_Z(y)\} \subseteq E_x \cup E_y, \qquad p_\psi(x, y) \leq \min\{1, \nu_R(x) + \nu_R(y)\} \leq c. \tag{36}$$

Coverage events at the two endpoints need not be independent. Applying either nondecreasing map $F_{T,b}$ or $G_{T,m}$ to the near and far bounds proves the corresponding branch of Theorem 3.3. □

The coverage probability follows directly from the sampling design. For iid references from μ, $\nu_R(x) = (1 - \mu(B(x, R/2)))^\psi$. For sampling without replacement, let m_x count the reference items within the closed ball. Then

$$\nu_R(x) = \frac{\binom{n-m_x}{\psi}}{\binom{n}{\psi}} \leq \left(1 - \frac{m_x}{n}\right)^\psi \leq e^{-\psi m_x/n}. \tag{37}$$

When the numerator is nonzero, the ratio is the product of factors $(n - m_x - i)/(n - i) \leq 1 - m_x/n$ for $i = 0, \ldots, \psi - 1$; otherwise the bound is immediate. Thus if each endpoint's ball contains at least a fraction $m > 0$ of reference items, a valid uniform far bound is $c = \min\{1, 2e^{-\psi m}\}$.

A.5 Amplification under a coupled signature budget

Let a, c be the single-round bounds from Theorem 3.3. For any $T = Lb$, elementary exponential and union bounds imply

$$\begin{aligned} \mathbb{P}[C^c \mid x, y \text{ fixed and near}] &\leq (1 - a^b)^L \leq e^{-La^b}, \\ \mathbb{P}[C \mid x, y \text{ fixed and far}] &\leq 1 - (1 - c^b)^L \leq Lc^b. \end{aligned} \tag{38}$$

Bounds larger than one are vacuous. In contrast to a second sketch applied to a fixed representation, adding a coordinate here means sampling another independent reference set.

For an existence guarantee, suppose $0 \leq c < a \leq 1$ and fix target errors $\xi_N, \xi_F \in (0, 1)$. Let $A = \log(1/\xi_N)$ and choose

$$L = \left\lceil \frac{A}{a^b} \right\rceil, \qquad T = bL. \tag{39}$$

The near miss rate is at most ξ_N, while

$$Lc^b \leq A(c/a)^b + c^b \longrightarrow 0 \quad \text{as } b \to \infty. \tag{40}$$

A sufficiently large integer b therefore also makes the far collision rate at most ξ_F. This may require a large T and hence a large encoding cost. At a fixed experimental budget such as $T = 100$, only the divisors of T are available as equal band lengths; the existence statement does not promise both target errors at that budget.

There are useful fixed-budget endpoint checks. Matching all T coordinates implies a candidate, and any candidate requires at least one matching coordinate. Hence

$$p^T \leq F_{T,b}(p) \leq 1 - (1 - p)^T. \tag{41}$$

The lower and upper endpoints are attained by $b = T$ and $b = 1$, respectively. Holding b fixed while adding complete independent bands increases capture for both near and far pairs. Holding T fixed while changing b trades band selectivity against the number of opportunities to match. Increasing ψ changes p_ψ itself and has no universal monotonic effect.

A.6 Proof of Corollary 3.4

Let $N = \{d(X,Y) \le r\}$, $H = \{d(X,Y) \le R\}$, and $F = H^c$. Integrating fixed-pair bounds over the specified pair distribution is valid because the pair draw is independent of encoder randomness. Since $N \subseteq H$,

$$u := \mathbb{P}[C \cap H] \ge \mathbb{P}[C \cap N] \ge \pi_N \alpha, \qquad v := \mathbb{P}[C \cap H^c] \le \pi_F \beta. \tag{42}$$

The assumption $\pi_N \alpha > 0$ ensures a nonzero conditioning probability. Since $u/(u+v)$ increases with u and decreases with v,

$$\mathbb{P}[H \mid C] = \frac{u}{u+v} \ge \frac{\pi_N \alpha}{\pi_N \alpha + \pi_F \beta}. \tag{43}$$

The intermediate region $H \setminus N$ only adds probability to the numerator event, so its mass need not be constrained. □

For $0 < \zeta < 1$, a sufficient condition for precision at least $1 - \zeta$ is

$$\pi_F \beta \le \frac{\zeta}{1-\zeta} \pi_N \alpha. \tag{44}$$

For a collision in one specified band, replace α, β by a^b, c^b. For a fixed weighted pair collection, one can also integrate pair-specific bounds. If w_i is a pair's probability, a_i lower-bounds its single-round collision when it is near, and c_i upper-bounds that probability when it is far, the same argument gives

$$\mathbb{P}[\text{near} \mid C] \ge \frac{\sum_{i \in N} w_i F_{T,b}(a_i)}{\sum_{i \in N} w_i F_{T,b}(a_i) + \sum_{i \in F} w_i F_{T,b}(c_i)}. \tag{45}$$

Here N, F partition the stated pair collection and the lower numerator is assumed positive. A zero lower numerator supplies only the trivial bound zero, whenever the conditioning event has positive probability. This version uses the actual margin at each pair's distance rather than the worst margin over the entire near population. These results concern a ratio of joint probabilities over the pair draw and encoder. They do not describe the expected precision of a uniformly sampled bucket or the precision of a particular realized index. Sampling queries and database items, sampling all unordered corpus pairs, and sampling buckets uniformly induce different pair priors.

A.7 Finite-precision nearest-index selection

The probability identities above concern exact distance ordering. The implementation evaluates the equivalent score $s(x,z) = \|z\|_2^2 - 2x^\top z$ using FP32 arithmetic; the omitted $\|x\|_2^2$ is constant across references. Near ties can therefore produce a different index than exact arithmetic. We quantify this effect without removing affected pairs from calibration.

Treat the stored FP32 coordinates as exact inputs, let $u = 2^{-24}$, and define $\gamma_k = ku/(1-ku)$ for $ku < 1$. Assume rounding to nearest, no overflow or underflow, and full FP32 multiplication. The standard product-and-sum expansion bounds the squared-norm error by $\gamma_d \sum_i z_i^2$ and the doubled dot-product error by $2\gamma_d \sum_i |x_i z_i|$. Rounding the final subtraction adds at most $u(1+\gamma_d)$ times their combined absolute scale. Thus

$$|\widehat{s}(x,z) - s(x,z)| \le \mathcal{E}(x,z) := \gamma_{d+3} \left(\|z\|_2^2 + 2 \sum_i |x_i z_i| \right), \tag{46}$$

where γ_{d+3} conservatively dominates $\gamma_d + u(1+\gamma_d)$. The bound concerns scores and makes no claim that rounded scores form a metric.

Let z_* minimize the exact score and let $\widehat{z}$ be the computed minimum. Since $\widehat{s}(x,\widehat{z}) \le \widehat{s}(x,z_*)$, any different selected index must satisfy

$$0 \le s(x,\widehat{z}) - s(x,z_*) \le \mathcal{E}(x,\widehat{z}) + \mathcal{E}(x,z_*). \tag{47}$$

Our CPU calibration checks this inequality for every observed FP32/FP64 index disagreement and stores both sets of candidate events. The two encoders disagree on 0.010% of nearest-index selections in the small-pool calibration of Appendix E.8. Every disagreement satisfies the bound, and none changes the reported complete-band events. The theoretical predictions use subtraction-based FP64 distance ordering as a numerical proxy for the ideal metric; the CPU FP32 encoder's events remain the primary empirical measurements. Agreement within Monte Carlo uncertainty does not establish bitwise equivalence of these two arithmetic paths.

A.8 Angular special case and assumption boundaries

An explicit population model connects the construction to angular hashing. Let $x, y \in \mathbb{S}^{d-1}$, $d \geq 2$, and sample two fresh iid uniform spherical references z_1, z_2 per round. Equal reference norms imply

$$\|x - z_1\|_2^2 \leq \|x - z_2\|_2^2 \quad \iff \quad (z_2 - z_1)^\top x \leq 0. \tag{48}$$

The direction $w = (z_2 - z_1)/\|z_2 - z_1\|_2$ is almost surely defined and uniform by rotational invariance. For any fixed nonzero input, ties have probability zero. If the angle between x, y is θ, the projected direction of w separates their signs on angular measure 2θ out of 2π. Thus

$$p_2(x, y) = 1 - \theta/\pi, \qquad \mathbb{P}[C(x, y)] = 1 - [1 - (1 - \theta/\pi)^b]^L. \tag{49}$$

This is the familiar random-hyperplane special case (Charikar, 2002), not a separate novelty claim. An arbitrary fixed finite reference collection need not have the rotational invariance required here.

Distance alone cannot guarantee locality for arbitrary reference collections. On the real line, take $D = \{0, 1\}$ and $\psi = 2$. For $0 < \epsilon < 1/2$, inputs $x = 1/2 - \epsilon$ and $y = 1/2 + \epsilon$ select distinct references in every round, so $p_\psi = F_{T,b}(p_\psi) = 0$ despite arbitrarily small separation. Conversely, $x = M, y = 2M$ for $M > 1$ always select reference 1, so the collision probability is one despite separation M.

The method name also does not make the classical set-MinHash identity automatic. For derived membership sets $S_T(x)$ and $S_T(y)$, M matching rounds do give Jaccard similarity $M/(2T - M)$, but the signature used here is H_T, not a uniform-permutation minimum of these sets. In particular, for a single band spanning all T coordinates, the direct collision probability is p_ψ^T. Substituting a Jaccard transformation into Equation (6) would describe a different algorithm.

A.9 Local evaluation with a prescribed collision error

Proposition A.1 (Collision probability from local neighborhoods)**.** *Let w_j be the contribution of reference j to Equation* (5)*, and set $q(k) = \binom{n-k}{\psi}/\binom{n}{\psi}$. Choose complete distance sublevel sets A around x and B around y, including all cutoff ties. With $I = A \cap B$, $U = A \cup B$, and $S_I = \sum_{j \in I} w_j$,*

$$S_I \leq p_\psi(x, y) \leq \min\{1, S_I + q(|A|) + q(|B|) - q(|U|)\}. \tag{50}$$

The competitor counts for $j \in I$ depend only on references in U; the binomial weights retain the original population size n. For $0 < \delta < 1$, if both neighborhoods contain at least k items with $2q(k) \leq \delta$, the interval width is at most δ.

We prove Proposition A.1 for a fixed pool of n indexed reference items and an ordered uniform sample Z of ψ distinct indices. All distances and parameters are fixed before the sample is drawn. Define the complete closed sublevel sets

$$A = \{j : d(x, z_j) \leq r_x\}, \quad B = \{j : d(y, z_j) \leq r_y\}, \quad I = A \cap B, \quad U = A \cup B.$$

Including every tied cutoff item is essential to the locality argument. Writing $k = |A|$, $l = |B|$, $i = |I|$, $u = |U|$, let

$$q(s) = \frac{\binom{n-s}{\psi}}{\binom{n}{\psi}}, \qquad w_j = \frac{q(a_j) - q(a_j + e_j + 1)}{e_j + 1}, \qquad S_I = \sum_{j \in I} w_j.$$

The finite-pool law identifies w_j with the probability that j is both inputs' winner, including the shared sampled-order tie rule.

Proof of Proposition A.1. The common-winner events are disjoint. If Z meets both A and B, a shared nearest reference lies in each complete sublevel set. Thus

$$\begin{aligned} 0 \leq p_\psi(x, y) - S_I &= \mathbb{P}[\text{a common winner exists outside } I] \\ &\leq \mathbb{P}[Z \cap A = \varnothing \text{ or } Z \cap B = \varnothing] = q(k) + q(l) - q(u) =: R. \end{aligned}$$

The last equality is inclusion–exclusion: missing both sets is missing their union. For $j \in I$ and $v \notin U$,

$$d(x, z_v) > r_x \geq d(x, z_j), \qquad d(y, z_v) > r_y \geq d(y, z_j).$$

Such an item is neither a strict nor a tied competitor of j. Computing (a_j, e_j) inside U therefore gives the global competitor counts, while n in q remains unchanged. Finally, if $k, l \geq k_0$ and $2q(k_0) \leq \delta$, then

$$0 \leq \min\{1, S_I + R\} - S_I \leq R \leq 2q(k_0) \leq \delta.$$

□

What local order information adds. The event with exactly one sampled union item, belonging to I, forces a common winner in I. Such a winner also requires sampling I. Therefore

$$\frac{i\binom{n-u}{\psi-1}}{\binom{n}{\psi}} \leq S_I \leq 1 - q(i). \tag{51}$$

A count-only lower bound uses the leftmost quantity; its corresponding upper bound is $\min\{1, 1 - q(i) + R\}$. The local-order interval refines both endpoints at the *same* cutoffs. It need not dominate a count-only interval optimized over other, larger neighborhoods. For identical rankings, multiple sampled intersection items still produce a shared winner; the local sum retains that probability mass.

Precision and computation. Choose the smallest k_0 satisfying the exact binomial condition $2\binom{n-k_0}{\psi} \leq \delta\binom{n}{\psi}$, and take each endpoint's complete k_0th-distance neighborhood. Since $q(k) \leq \exp(-\psi k/n)$, the value

$$\min\left\{n, \left\lceil \frac{n}{\psi} \log \frac{2}{\delta} \right\rceil\right\}$$

is also sufficient. Without cutoff ties $u \leq \min\{n, 2k_0\}$; with ties we report the actual u, which can equal n. Given full distances, selection and membership take $O(n)$ work, followed by $O(u \log u)$ local dominance counting. Direct distance acquisition still costs $O(nd)$, and exact binomial evaluation has its own arithmetic cost. Full-pool dominance counts can be reused across all ψ values; separate local neighborhoods may require repeated counts. A smaller counting domain for one setting therefore does not establish faster complete parameter selection. This offline diagnostic leaves the retrieval algorithm unchanged.

Propagation to candidate probabilities. For $p \leq p'$, use the same T iid uniform variables to couple the Bernoulli agreements at both probabilities. The band and count events are monotone. They can differ only if some uniform lies in $(p, p']$, so for $\Phi \in \{F_{T,b}, G_{T,m}\}$,

$$0 \leq \Phi(p') - \Phi(p) \leq \mathbb{P}[\exists t : V_t \in (p, p']] \leq T(p' - p). \tag{52}$$

Applying each gate to the interval endpoints consequently gives a candidate-probability interval of width at most $T\delta$. Choosing $\delta = \varepsilon/T$ controls that width by ε. Endpoints must be composed pairwise before averaging. These statements condition on a fixed pool and pair; they are neither a certificate for an adaptively selected fitted index nor a population precision guarantee.

A.10 An exact locality curve on the unit circle

The iid population model admits a simple case in which both forward collision probabilities and reverse proximity probabilities are explicit. Let the inputs lie on the unit circle, and let each round draw $\psi \geq 2$ references independently and uniformly from the circle. All inputs share the ordered references within a round, and different rounds draw fresh references independently. The output is the local nearest index $h_Z(x) = \arg\min_{0 \leq j < \psi} \|x - z_j\|_2$. Euclidean chord distance, angular distance, and cosine distance induce the same ordering on this domain; ties have probability zero.

Proposition A.2 (Uniform-circle collision probability)**.** *For a fixed input pair with shortest angle $\theta \in [0, \pi]$, write $s = \theta/\pi$. Its single-round collision probability is*

$$p_\psi(\theta) = (1 - s)^{\psi - 1} \left(1 + \frac{\psi - 2}{2} s\right). \tag{53}$$

Proof. Start the cyclic ordering at one labeled reference. Rotating that reference to angle zero leaves the other $\psi - 1$ reference angles iid uniform, so the ψ normalized circular gaps have the $\mathrm{Dirichlet}(1, \ldots, 1)$ distribution. The arc on which this reference is nearest has length ℓ equal to half

the sum of its two adjacent gaps. For $\psi \geq 3$, therefore, $U = \ell/\pi \sim \mathrm{Beta}(2, \psi - 2)$, and $\ell \leq \pi$. For an arc of length ℓ, the starting angles for which both endpoints of a pair separated by $\theta \leq \pi$ lie in that arc have total length $(\ell - \theta)_+$. Rotational invariance and exchangeability of reference labels give

$$p_\psi(\theta) = \frac{\psi}{2\pi}\mathbb{E}[(\ell - \theta)_+] = \frac{\psi}{2}\int_s^1 \mathbb{P}[U > u]\, du. \tag{54}$$

With $m = \psi - 2$, the Beta survival function is $\mathbb{P}[U > u] = (1-u)^m(1+mu)$. Hence

$$p_\psi(\theta) = \frac{\psi}{2}\left[(1-s)^{m+1} - \frac{m}{m+2}(1-s)^{m+2}\right], \tag{55}$$

which simplifies to Equation (53). For $\psi = 2$, both nearest-reference arcs are semicircles, yielding $p_2(\theta) = 1 - s$ directly. □

Forward locality and direct banding. The curve decreases strictly for $0 \leq \theta < \pi$:

$$\frac{dp_\psi}{d\theta} = -\frac{\psi}{2\pi}(1-s)^{\psi-2}[1 + (\psi-2)s] < 0, \qquad p_\psi(0) = 1, \quad p_\psi(\pi) = 0. \tag{56}$$

For $T = Lb$ independent rounds and complete bands of length b, Proposition 3.1 gives $\mathbb{P}[C(x,y)] = F_{T,b}(p_\psi(\theta))$. Thus, for angular thresholds $0 < r < R < \pi$,

$$\theta \leq r \implies \mathbb{P}[C(x,y)] \geq F_{T,b}(p_\psi(r)), \qquad \theta \geq R \implies \mathbb{P}[C(x,y)] \leq F_{T,b}(p_\psi(R)). \tag{57}$$

These bounds require no unspecified margin or coverage parameters in this population model. Angular thresholds translate to Euclidean thresholds through $d = 2\sin(\theta/2)$.

Reverse proximity under a specified pair prior. Draw X, Y independently and uniformly from the circle, independently of the reference samples, and let Θ be their shortest angle. Then Θ/π is uniform on $[0,1]$. For the single-round event $C_1 = \{h_Z(X) = h_Z(Y)\}$, integrating the collision curve gives

$$\begin{aligned} \mathbb{P}[C_1] &= \int_0^1 p_\psi(\pi u)\, du = \frac{3}{2(\psi+1)}, \\ \mathbb{P}[\Theta \leq r \mid C_1] &= 1 - (1-s)^\psi\left(1 + \frac{\psi-2}{3}s\right), \qquad s = r/\pi \in [0,1]. \end{aligned} \tag{58}$$

Indeed, the tail integral is $\int_s^1 p_\psi(\pi u)\, du = (1-s)^\psi[3 + (\psi-2)s]/[2(\psi+1)]$; normalizing it by $\mathbb{P}[C_1]$ proves the second identity. For the complete band candidate event, the exact posterior is

$$\mathbb{P}[\Theta \leq r \mid C] = \frac{\int_0^{r/\pi} F_{T,b}(p_\psi(\pi u))\, du}{\int_0^1 F_{T,b}(p_\psi(\pi u))\, du}. \tag{59}$$

These probabilities average over the pair draw and reference sampling. Conditioning a uniformly sampled pair on collision weights a single-round region by its squared arc length; uniformly selecting a region and then drawing two points inside it defines a different prior.

For example, take $\psi = 32, T = 100, b = 5$. Evaluating the preceding expressions yields the conservative bounds

$$\theta \leq 0.02\pi \implies \mathbb{P}[C(x,y)] \geq 0.971, \qquad \theta \geq 0.25\pi \implies \mathbb{P}[C(x,y)] \leq 2.085 \times 10^{-15}. \tag{60}$$

Under the uniform pair prior, the probability of $\Theta \leq 0.05\pi$ is 0.05, whereas

$$\mathbb{P}[\Theta \leq 0.05\pi \mid C] \simeq 0.975. \tag{61}$$

These are analytic evaluations of the population model.

Conditioning on a finite reference collection. This special case uses fresh iid population references. The finite-collection results of Appendix A instead condition on D and use its collision probability $p_{\psi,D}(x,y)$. If D itself consists of $n \geq \psi$ iid uniform circle points, x, y are fixed independently of D, and a round samples without replacement, then $\mathbb{E}_D[p_{\psi,D}(x,y)] = p_\psi(\theta)$. However, multiple rounds share D, so marginalizing it gives

$$\mathbb{P}[C(x,y)] = \mathbb{E}_D[F_{T,b}(p_{\psi,D}(x,y))], \tag{62}$$

which generally differs from $F_{T,b}(p_\psi(\theta))$. The marginal single-round identity need not hold when the input pair is selected from D itself. Consequently, this example establishes a nontrivial two-direction locality result under its stated population and pair assumptions; it does not replace the finite-collection collision law or certify every realized bucket.

A.11 COLLISION-COUNTING CANDIDATE SELECTION

Collision-counting LSH (Gan et al., 2012) retains a pair when enough coordinates agree, wherever those agreements occur. Applied directly to SMinHash's nearest-reference indices, it preserves the encoder and generally accepts different patterns from disjoint banding.

A.11.1 FIXED-PARAMETER PROBABILITIES AND LOCALITY

Fix the reference collection D of size n, parameters $1 \le \psi \le n$, $T \ge 1$, and an integer threshold $1 \le m \le T$. Rounds independently draw shared ordered samples of ψ distinct items from D. The nearest-reference coordinate h_t uses the first local index on ties. Define

$$S_T(x,y) = \sum_{t=1}^{T} \mathbf{1}\{h_t(x) = h_t(y)\}, \qquad A_m(x,y) = \{S_T(x,y) \ge m\}. \tag{63}$$

All probabilities condition on D.

Proposition A.3 (Count acceptance and separated pairs)**.** *For a fixed pair, let $p = p_{\psi,D}(x,y) = \mathbb{P}[h_t(x) = h_t(y)]$. Then $S_T \sim \mathrm{Binomial}(T,p)$ and*

$$\mathbb{P}[A_m(x,y)] = G_{T,m}(p), \qquad G_{T,m}(p) = \sum_{j=m}^{T} \binom{T}{j} p^j (1-p)^{T-j}. \tag{64}$$

The endpoints are interpreted by continuity. Suppose, on specified query/database domains, all pairs in $\mathcal{N} = \{(x,y) : d(x,y) \le r\}$ satisfy $p \ge p_1$, and all pairs in $\mathcal{F} = \{(x,y) : d(x,y) > R\}$ satisfy $p \le p_2$, where $0 < r \le R$ and $0 \le p_2 < p_1 \le 1$. Put $a_m = G_{T,m}(p_1)$ and $c_m = G_{T,m}(p_2)$. Then

$$(x,y) \in \mathcal{N} \implies \mathbb{P}[A_m(x,y)] \ge a_m, \qquad (x,y) \in \mathcal{F} \implies \mathbb{P}[A_m(x,y)] \le c_m. \tag{65}$$

If $p_2 < \tau = m/T < p_1$, the respective errors obey

$$\sup_{\mathcal{N}} \mathbb{P}[A_m^c] \le e^{-2T(p_1-\tau)^2}, \qquad \sup_{\mathcal{F}} \mathbb{P}[A_m] \le e^{-2T(\tau-p_2)^2}. \tag{66}$$

Proof. The agreement indicators are iid Bernoulli variables with parameter p, giving Equation (64). Differentiating its finite sum and canceling adjacent terms yields

$$G'_{T,m}(p) = T\binom{T-1}{m-1} p^{m-1}(1-p)^{T-m} \ge 0, \qquad 0 < p < 1. \tag{67}$$

Monotonicity proves Equation (65). For near pairs, A_m^c implies $S_T/T < \tau < p_1 \le p$; for far pairs, A_m implies $S_T/T \ge \tau > p_2 \ge p$. Applying the one-sided Hoeffding inequality at deviations $p_1 - \tau$ and $\tau - p_2$ proves Equation (66). □

Theorem 3.3 supplies p_1, p_2 when its geometric bounds are separated; S_T/T estimates this collision probability. The endpoints $m = 1, T$ recover singleton-band OR and one complete-signature band. For $T = 4$, patterns 1010 and 1100 both pass $m = 2$, whereas only the latter matches a length-two band.

Reverse proximity requires a pair prior. Draw (X,Y) independently of the reference rounds, with near, far, and intermediate prior masses π_N, π_F, π_G, summing to one. If $\pi_N a_m > 0$, then

$$\mathbb{P}[(X,Y) \in \mathcal{N} \mid A_m] \ge \frac{\pi_N a_m}{\pi_N a_m + \pi_F c_m + \pi_G}, \mathbb{P}[d(X,Y) \le R \mid A_m] \ge \frac{\pi_N a_m}{\pi_N a_m + \pi_F c_m}. \tag{68}$$

To prove this, write u_N, u_F, u_G for the probabilities of acceptance together with each class. Equation (65) gives $u_N \ge \pi_N a_m$, $u_F \le \pi_F c_m$, and $0 \le u_G \le \pi_G$. The first posterior is $u_N/(u_N + u_F + u_G)$; the second is $(u_N + u_G)/(u_N + u_F + u_G)$. Their monotonicity in these masses gives the two bounds. They average over the specified pair draw and random index, and do not certify every realized bucket or every accepted pair.

A.11.2 A LOSSLESS MULTI-INDEX PREFILTER

Bucket lookups followed by complete signature verification implement the same predicate through the pigeonhole principle of multi-index hashing (Norouzi et al., 2012, Section 2).

Proposition A.4 (Complete recovery of the count candidates)**.** *Let $L_c = T - m + 1$ and partition the T coordinate positions into L_c disjoint nonempty groups $I_1, \ldots, I_{L_c}$. Index each group separately by its subtuple. Let $B(x, y)$ denote equality of at least one group subtuple. For every realized signature pair, including distance ties,*

$$A_m(x, y) \implies B(x, y), \qquad B(x, y) \cap \{S_T(x, y) \geq m\} = A_m(x, y). \tag{69}$$

Thus unioning the matching group buckets and verifying the original T coordinates recovers exactly the count candidates.

Proof. If every group fails to match, each contains a mismatch. Disjointness gives at least L_c distinct mismatches, so

$$B(x, y)^c \implies T - S_T(x, y) \geq L_c \implies S_T(x, y) \leq m - 1. \tag{70}$$

The contrapositive proves the inclusion; intersection with the full count predicate proves exact recovery. This argument is deterministic and requires no independence or distributional assumption. □

Balanced groups have sizes $g_\ell \in \{\lfloor T/L_c \rfloor, \lceil T/L_c \rceil\}$. Shorter groups may be padded to a common key width with constant zeros, provided group identities stay separate and padding is excluded from S_T. Empty groups are unnecessary and uninformative. The prefilter preserves any predeclared self-exclusion or other eligibility rule when it is applied consistently to both candidate sets.

For a partition fixed before the reference draws, independence of the rounds gives the exact prefilter probability

$$u(p) = 1 - \prod_{\ell=1}^{L_c} (1 - p^{g_\ell}), \qquad G_{T,m}(p) \leq u(p). \tag{71}$$

Indeed, group events involve disjoint sets of independent indicators; the inequality also follows from Equation (69). For a fixed query and eligible database items y_i, set $p_i = p_{\psi,D}(x, y_i)$. If W is the number of posting visits, U the prefilter union size, and C the final count-candidate size, then

$$\mathbb{E}W = \sum_{i=1}^{N} \sum_{\ell=1}^{L_c} p_i^{g_\ell}, \qquad \mathbb{E}U = \sum_{i=1}^{N} u(p_i), \qquad \mathbb{E}C = \sum_{i=1}^{N} G_{T,m}(p_i). \tag{72}$$

These expectations require no independence between database items. The full agreement total has expectation $T \sum_i p_i$; counting only prefilter items gives a different total. Small $p_i^{g_\ell}$ can yield sparse unions, while many short groups can make $u(p_i)$ close to one. Alphabet size alone determines neither regime.

Work and memory. After encoding, group-key formation reads NT coordinates and the index stores NL_c postings, alongside the full signatures. With expected constant-time bucket lookup and deduplication per visit, one query takes $O(T+L_c+W+UT)$ work through count verification; scoring all C survivors in d dimensions adds $O(Cd)$, before the final top-k reduction. Direct encoding itself costs $O(NT\psi d)$ for the database and $O(T\psi d)$ per query in the unreused implementation. The index requires $O(NT + NL_c)$ coordinate/ID words; a dense batched counter or a candidate workspace allocated for every possible query–database pair additionally requires $O(QN)$ words of scratch, even when its active prefix is small. These costs do not imply universally sublinear retrieval or linear-time graph construction: U and C can equal N, and an all-pairs output can be quadratic. No candidate cutoff is part of Equation (63).

A.11.3 WHAT REMAINS VALID AFTER ADAPTIVE SELECTION?

Equation (64) fixes parameters before drawing the index. A validation-selected $\widehat{m}$ can share reference randomness with $S_T(x, y)$ even for independent validation and evaluation examples. Thus conditioning on $\widehat{m} = m$ need not recover $G_{T,m}(p)$. Held-out evaluation remains valid as an empirical protocol, while conditioning on the whole fitted index makes a fixed pair's event deterministic.

Two bounds survive such dependence. First, for deterministic endpoints $m_- \le \widehat{m} \le m_+$ and fixed ψ, T,

$$\{S_T \ge m_+\} \subseteq \{S_T \ge \widehat{m}\} \subseteq \{S_T \ge m_-\} \quad \Longrightarrow \quad G_{T,m_+}(p) \le \mathbb{P}[A_{\widehat{m}}] \le G_{T,m_-}(p). \tag{73}$$

Second, declare M input pairs independently of the reference draws and J candidate encoder families in advance. Family j has T_j iid rounds and pair probability p_{ji}. Hoeffding and a union bound give, for any deterministic $\epsilon_j > 0$,

$$\mathbb{P}[\exists j, i : |S_{ji}/T_j - p_{ji}| > \epsilon_j] \le 2M \sum_{j=1}^{J} e^{-2T_j\epsilon_j^2}. \tag{74}$$

No independence between pairs or between families is required. On the complementary event, the inequalities hold after any choice of family and threshold, simultaneously over all declared pairs; in particular, acceptance implies $p_{ji} \ge m/T_j - \epsilon_j$. There is no additional union over thresholds. This finite-pair statement does not extend to arbitrary pairs chosen after inspecting the index. Alternatively, parameters chosen using an independent pilot index recover the fixed-parameter law when fresh reference rounds are drawn. Finally, adapting the prefilter partition does not affect its deterministic completeness, although it can invalidate the fixed-partition work probabilities in Equations (71)–(72).

B Bucket populations and candidate work

Label symmetry and bucket balance use different probability spaces. The former varies the reference sample for a fixed input; the latter varies inputs after fixing an index. Entropy and bit-independence arguments for encoded reference labels appear in Xu & Ting (2025, Sec. 4.4) and Zhang et al. (2026, App. C). Here we make the conditioning explicit for the first-local-index rule, including ties, and connect the pairwise collision law to candidate work. The dependence of query work on squared bucket sizes already appears in Kang & Jung (2012, Secs. 1, 3). The elementary identities below specify directed non-self counts, repeated band matches and conditioning for the present rule; they do not introduce another encoder.

B.1 The marginal label law with distance ties

Fix an input x and an unordered sample of ψ indexed references. Let M be the number attaining the smallest distance to x. Only the uniform ordering of this sample is random in the next statement.

Proposition B.1 (First minimum under a random order)**.** *For $1 \le m \le \psi$ and $0 \le j < \psi$,*

$$\mathbb{P}[h(x) = j \mid M = m] = \frac{\binom{\psi-j-1}{m-1}}{\binom{\psi}{m}}. \tag{75}$$

After also averaging over unordered samples, $\mathbb{P}[h(x) = 0] = \mathbb{E}[M]/\psi$. Thus the marginal label is uniform if and only if the nearest reference is unique almost surely under the sampling law.

Proof. The m minimizing references occupy a uniformly selected m-subset of the ψ positions. The first occupied position is j exactly when j is selected and the remaining $m-1$ positions lie after it. Hence

$$\#\{\text{minimum-position sets starting at } j\} = \binom{\psi-j-1}{m-1}, \qquad \mathbb{P}[h(x) = 0 \mid M = m] = \frac{m}{\psi}.$$

For $M = 1$ the first expression is $1/\psi$ for every position. Conversely, uniform labels imply $\mathbb{E}[M] = 1$; since $M \ge 1$, this forces $M = 1$ almost surely. □

When $\psi = 2^w$ and the minimum is unique almost surely, the uniform label has w independent fair binary digits, over encoder randomness for this fixed input. With ties this conclusion can fail: for $\psi = 4$ and $M = 2$, the probabilities of labels $00, 01, 10, 11$ are $(1/2, 1/3, 1/6, 0)$. The probability of two one bits is zero, whereas the product of their marginals is $1/18$. This example shows possible dependence, not that every tied case has dependent bits. Independent reference rounds still give independent coordinate agreements for a fixed pair, so ties do not invalidate the band and count laws proved in Appendix A.3.

B.2 From collision probabilities to candidate work

Fix N indexed database items and a fitted index with L bands of length b. Let $n_{\ell g}$ be the number of items with key g in band ℓ. Let W count directed non-self pair occurrences across all bands, including repetitions, and let C_N count distinct directed non-self candidate pairs after their union. Then

$$W = \sum_{\ell=1}^{L} \sum_{g} n_{\ell g}(n_{\ell g} - 1), \qquad \frac{W}{L} \le C_N \le W. \tag{76}$$

The first equality counts the ordered distinct-index pairs in each bucket. Each admitted pair contributes between one and L occurrences, which proves the inequalities. These are logical pair counts; bitsets can avoid explicitly visiting every repeated occurrence.

Now fix the reference collection and database vectors, and average only over independent encoder rounds. Writing $p_{ij} = p_\psi(x_i, x_j)$ gives

$$\mathbb{E}[W] = L \sum_{i \neq j} p_{ij}^b, \qquad \mathbb{E}[C_N] = \sum_{i \neq j} F_{T,b}(p_{ij}). \tag{77}$$

Indeed, a specified pair matches a specified band with probability p_{ij}^b and passes the union with probability $F_{T,b}(p_{ij})$. Linearity of expectation requires no independence between pairs. For complete count admission, the expected distinct candidate count is instead $\sum_{i \neq j} G_{T,m}(p_{ij})$.

The geometric bounds also give a sufficient workload condition. Let M_R count the directed pairs with distance at most R. If every other pair has acceptance probability at most β, then

$$\mathbb{E}[C_N] \le M_R + \{N(N-1) - M_R\}\beta. \tag{78}$$

Thus $M_R = O(N)$ and $\beta = O(1/N)$ suffice for $O(N)$ expected distinct candidates. They do not follow from a fixed signature length or from label symmetry. Encoding, indexing and verification costs remain separate; this is not an unconditional linear-time graph result.

B.3 Why uniform labels do not imply balanced buckets

Let X, Y be iid from a fixed input distribution, independently of the random encoder. For one fitted band B, write $q_g = \mathbb{P}_X[B(X) = g \mid B]$. Conditional independence of the two inputs gives

$$\mathbb{P}[B(X) = B(Y) \mid B] = \sum_{g} q_g^2. \tag{79}$$

Suppose the nearest reference is unique almost surely for almost every input. A fixed input's band key is then uniform over $K = \psi^b$ keys under encoder randomness. Consequently $\mathbb{E}_B[q_g] = 1/K$, but

$$\mathbb{E}_B\left[\sum_{g} q_g^2\right] = \frac{1}{K} + \sum_{g} \mathrm{Var}_B(q_g). \tag{80}$$

This follows by expanding each second moment. Uniform marginal labels therefore give a lower bound $1/K$ on random-pair collision, not an upper bound or balanced realized buckets. For the empirical distribution of N database items, $q_g = n_g/N$; removing identical item indices changes the one-band pair count from $N^2 \sum_g q_g^2$ to $N^2 \sum_g q_g^2 - N$, agreeing with Equation (76).

B.4 A rank-neighborhood certificate for the maximum load

All database/reference distances in this argument are fixed before the reference samples. The neighborhoods in Equation (9) contain all items tied at the kth cutoff. Let U_t count the database items whose neighborhood is missed in round t.

Proof of Proposition 3.2. If the sample meets $\mathcal{N}_k(x)$, its nearest reference is also in that set. Thus at most R_k covered inputs choose any one reference; all remaining inputs number at most U_t. Writing n_{tj} for a coordinate bucket's population, deterministically,

$$\max_j n_{tj} \le R_k + U_t, \qquad U_t = \sum_{x \in \mathcal{X}} \mathbf{1}\{Z_t \cap \mathcal{N}_k(x) = \varnothing\}. \tag{81}$$

Uniform sampling without replacement gives

$$\mathbb{E}[U_t] = \sum_{x \in \mathcal{X}} \frac{\binom{n-|\mathcal{N}_k(x)|}{\psi}}{\binom{n}{\psi}} \leq N q_k. \tag{82}$$

The inequality uses $|\mathcal{N}_k(x)| \geq k$; no independence between the indicators is assumed. Markov's inequality and a union bound yield

$$\mathbb{P}[M_* \geq R_k + s] \leq \sum_{t=1}^{T} \mathbb{P}[U_t \geq s] \leq \frac{1}{s} \sum_{t=1}^{T} \mathbb{E}[U_t] \leq TNq_k/s.$$

Capping a probability at one proves the statement. Independence across rounds is not required for this particular bound. □

Choosing a bound before fitting the index. For a failure probability $\delta \in (0, 1]$, define

$$\Lambda_k(\delta) = \min\left\{ N,\ R_k + \max\left(0, \left\lceil \frac{TNq_k}{\delta} \right\rceil - 1\right)\right\}. \tag{83}$$

Then $\mathbb{P}[M_* > \Lambda_k(\delta)] \leq \delta$. For $q_k = 0$ every neighborhood is hit almost surely. Otherwise strict exceedance requires $U_t \geq \lceil TNq_k/\delta \rceil$ in some round, giving the claim by the preceding proof. The cap at N is deterministic.

For any fixed finite rank grid, the minimizer of Λ_k depends only on the fixed scores and (ψ, T, δ). It can therefore be chosen without a union bound over ranks. Choosing a rank from observed fitted bucket sizes would not have this justification. Simultaneous claims across datasets, resolutions and fits need separate multiplicity control, as in Figure 2.

Every band bucket is contained in each constituent coordinate bucket. On $M_* \leq \Lambda_k$, the pair-occurrence count of Equation (76) satisfies $W \leq LN(\Lambda_k - 1)$. A large Λ_k can still give quadratic work; this is not an unconditional linear graph guarantee.

Including self entries, complete self-query postings satisfy $P = LN + W \leq LN\Lambda_k$ on the same event. Across the four datasets, ten reference resolutions and eight band widths in the task grid, this consequence certifies none of the 320 conditions against the 10^9 posting budget, either at the primary or simultaneous confidence level. This is a limitation of the sufficient bound, not evidence that all 320 fitted indexes exceed the budget. An informative coordinate-load bound need not certify affordable complete candidate processing.

Scope and computation. The proof also holds for a fixed score matrix when its rankings and first-local-minimum evaluations use the same entries. It does not require a triangle inequality, and the numerical study uses this fixed-score interpretation. Its certificate consequently does not bound buckets from a different floating-point evaluation path. Computing every R_k requires database/reference score acquisition and rank information. Streaming reduces storage, not that computation; the certificate is an offline diagnostic rather than a free index check.

For metric distances, with $r = \max_x r_k(x)$, the set of inputs having $z \in \mathcal{N}_k(x)$ lies in $B(z, r)$. Thus $R_k \leq M_r := \max_{z \in D} |\mathcal{X} \cap B(z, r)|$ at this particular common radius. The rank certificate permits missed neighborhoods; it does not require every local ball to be sampled.

The empirical rank grid contains powers of two through n and $3 \cdot 2^j$ for $j = 5, \ldots, 12$. Within each of three independent fits, ordered samples of 1,024 distinct reference IDs supply uniform prefixes for all ten ψ values, with $T = 100$ rounds. The primary bound uses $\delta = 0.05$ for each dataset/resolution/fit; $\delta = 0.05/120$ gives a joint failure probability at most 0.05 over the whole diagnostic. Means and sample standard deviations summarize the observed maxima; three fits do not independently establish the advertised tail probability.

B.5 Coverage does not imply balanced loads

A region containing m of N possible references receives $\psi m/N$ references in expectation and is missed with probability $\binom{N-m}{\psi}/\binom{N}{\psi} \leq e^{-\psi m/N}$. When the reference pool is the entire database

of distinct points in a metric space, every sampled point chooses itself. All ψ coordinate buckets are therefore nonempty, and their largest load is at most $N - \psi + 1$. Neither property makes this maximum close to N/ψ.

Geometry in a fixed dimension. In $\mathbb{R}^{d+1}$, $d \geq 2$, let the database contain H hub points h and $B = N - H$ bulk points x, where $1 \leq H < N$. All points have unit norm, with

$$x = (a_x, \sqrt{1 - a_x^2}\, u_x), \quad u_x \in \mathbb{S}^{d-1}, \quad a_x \in [a_-, a_+], \quad \|h - e_0\| \leq \eta,$$
$$0 < a_- \leq a_+ < 1, \quad 0 \leq 3\eta < \sqrt{2 - 2a_+}, \qquad \rho = \frac{a_- - a_+^2 - \eta}{1 - a_-^2} \in (0, 1). \tag{84}$$

Define the largest directional-cap population, including its center,

$$C_D = \max_{v \in D_{\text{bulk}}} |\{x \in D_{\text{bulk}} : u_x^\top u_v \geq \rho\}|. \tag{85}$$

Every round samples ψ distinct references from the complete database, in a uniform order, and applies the unchanged nearest-index rule.

Proposition B.2 (A shared reference can create a large bucket)**.** *Under Equation* (84)*, fix D and set*

$$e_\psi = \frac{H\binom{B}{\psi-1}}{\binom{N}{\psi}}, \qquad L_{\psi,b}(c) = H + \max\{0, B - b(\psi - 1)c\}. \tag{86}$$

For T independent rounds, the largest coordinate bucket M_* satisfies

$$\mathbb{P}[M_* \geq L_{\psi,1}(C_D) \mid D] \geq 1 - (1 - e_\psi)^T. \tag{87}$$

For $b \mid T$, the largest bucket over the T/b disjoint bands has size at least $L_{\psi,b}(C_D)$ with probability at least $1 - (1 - e_\psi^b)^{T/b}$.

Proof. For any bulk input x, hub h and bulk reference v with $u_x^\top u_v < \rho$,

$$x^\top h \geq a_x - \eta \geq a_- - \eta, \qquad x^\top v < a_+^2 + (1 - a_-^2)\rho = a_- - \eta.$$

For a negative directional inner product, the second bound follows by replacing its contribution by zero; otherwise each square-root factor is at most $\sqrt{1 - a_-^2}$. Unit norms turn these inequalities into a strict distance preference for h. Each bulk reference can therefore compete with a hub for at most C_D inputs. All hub inputs also prefer any hub reference: hub–hub distances are at most 2η, whereas hub–bulk distances are at least $\sqrt{2 - 2a_+} - \eta > 2\eta$.

A sample contains exactly one hub with probability e_ψ. On that event, the hub wins for all H hub inputs and all but at most $(\psi - 1)C_D$ bulk inputs, proving the single-round load bound. Independent rounds give Equation (87). If all b rounds of a band contain exactly one hub, the intersection of their hub buckets contains at least $L_{\psi,b}(C_D)$ points. They share the entire tuple even if the winning local indices differ between rounds. Independence across disjoint bands gives the final bound. □

Continuous distributions with controlled cap population. Take $a_- < a_+$ and $\eta > 0$. Draw the bulk directions iid uniformly on $\mathbb{S}^{d-1}$ and their first coordinates independently from a density on $[a_-, a_+]$. Independently draw hubs from a continuous spherical cap satisfying Equation (84). For $\eta > 0$, uniform first coordinate on $[1 - \eta^2/2, 1]$ and uniform remaining direction is one such distribution. Both components have densities with respect to spherical surface measure; points and nearest-reference minima are distinct almost surely. Fixed component counts also describe an iid mixture conditional on those counts.

For a fixed direction v, rotational invariance gives

$$s_d(\rho) = \mathbb{P}[U^\top v \geq \rho] = \frac{\int_\rho^1 (1 - z^2)^{(d-3)/2}\, dz}{\int_{-1}^1 (1 - z^2)^{(d-3)/2}\, dz} = \tfrac{1}{2} I_{1-\rho^2}\left(\tfrac{d-1}{2}, \tfrac{1}{2}\right), \tag{88}$$

where I is the regularized incomplete beta function. Conditional on a bulk direction, its cap population is $1 + \text{Bin}(B - 1, s_d)$. A union bound over centers yields, for any integer $c \geq 1$,

$$\mathbb{P}[C_D > c] \leq \delta_c := \min\{1, B\mathbb{P}[\text{Bin}(B - 1, s_d) \geq c]\}. \tag{89}$$

No independence between different cap populations is required. Since e_ψ depends only on the component counts, conditioning first on D and then averaging proves the joint data-and-encoder guarantee

$$\mathbb{P}[M_* \geq L_{\psi,1}(c)] \geq (1-\delta_c)\{1-(1-e_\psi)^T\}. \tag{90}$$

The band analogue follows in the same way. This is a finite-N result at fixed dimension; it does not assert a limit uniform in N, since the cap population increases with B.

Numerical instance and paired diagnostic. Take $N = 65{,}536$, $H = 64$, $d = 256$, $a_- = 0.499$, $a_+ = 0.501$, $\eta = 0.005$, and $T = 100$. Numerical evaluation gives $s_d \approx 5.624 \times 10^{-8}$ and $\delta_3 \approx 5.433 \times 10^{-4}$. At $\psi = 1024$, $e_\psi \approx 0.371$ and $L_{\psi,1}(3)/N \approx 0.953$. Thus large loads do not require repeated points, distance ties or dimension growing with the database.

We evaluate the construction using three independent encoder fits on one fixed database, paired with a control that replaces the 64 hubs by ordinary bulk points. Reference IDs are coupled across arms and the ten $\psi \in \{2, 4, \ldots, 1024\}$ settings. FP64 evaluation and independent CPU checks retain the same direct minimum-index rule. At $\psi = 1024$, the mean maximum coordinate load is 98.439% with hubs and 0.156% without them; sample standard deviations are 0.000 and 0.003 percentage points. The corresponding five-row band maxima are $34.541\% \pm 21.257$ and $0.002\% \pm 0.001$ percentage points. At this resolution, each treated coordinate maximum attains $N - \psi + 1$, despite all buckets being nonempty. These are construction checks, not performance comparisons or empirical estimates of the probability in Equation (90).

Sparse true output can still produce quadratic candidates. A separate family makes the candidate-work implication exact. Let

$$a_i = \frac{1}{i+1}, \quad x_i = a_i e_0 + \sqrt{1-a_i^2}\, e_i, \quad D = \mathcal{X}_N = \{x_i\}_{i=1}^N \subset \mathbb{R}^{N+1}. \tag{91}$$

For $i \neq j$, $\|x_i - x_j\|^2 = 2 - 2a_i a_j$. Sampled items choose themselves, while every unsampled item uniquely chooses the sampled reference with smallest index. Each round therefore has one bucket of size $N - \psi + 1$ and $\psi - 1$ singletons, with directed-pair collision fraction

$$\frac{(N-\psi+1)(N-\psi)}{N(N-1)}. \tag{92}$$

In a band, at least $q = \max\{0, N - b\psi\}$ items occur in none of its samples and share its complete key, so

$$C_N \geq q(q-1). \tag{93}$$

For $N \geq 2$, the threshold $\sqrt{41/24}$ has only the two directed edges between x_1, x_2: their squared distance is $5/3$, while every other distinct pair has squared distance at least $7/4$. This proves quadratic candidate volume at fixed b, ψ despite sparse true output. Its dimension grows with N, unlike the preceding continuous construction. Both examples retain uniform fixed-input labels over random reference orders, which therefore cannot certify fitted loads.

C Datasets and evaluation protocols

How the experiments fit together. The primary reference-source comparison independently tunes both sources and evaluates new query IDs (Appendix E.1). Earlier banded studies evaluate retrieval, semantic screening, graphs and joins. The eight-dataset count study tests candidate quality on a separate new-query sample (Appendix D.1); its subsequent complete-retrieval comparison reuses those queries and reports returned IDs and timing (Appendix F.2). Request-size and execution ablations use development queries and separate configurations (Appendices F.3 and F.4). These studies answer different questions and their timing or quality measurements are not pooled.

C.1 Collision calibration

The calibration fixes 64 reference vectors and 16 pairs on each of Sphere64, Mixture64, GloVe25, and COCO512. Eight pairs come from a nearest-pool population and eight from uniform pool draws. For each $\psi \in \{2, 8, 32\}$ we fit 1,024 independent encoders with 100 independently sampled,

uniformly ordered reference rounds. Exact finite-pool probabilities include sampled-order ties as in Appendix A.1.

We evaluate every complete band width dividing $T = 100$, including a single band of width 100, and prefix controls $(T, b) = (25, 5)$ and $(50, 5)$. The main calibration fixes $(T, b) = (100, 5)$. Uncertainty is measured over complete independently fitted encoders, preserving dependence among pairs sharing an encoder. These CPU probability checks do not measure retrieval runtime. Compared with subtraction-based FP64 distances, 0.010% of FP32 nearest indices differ, all within the score-error envelope. Every pair is retained. The reported complete band events agree across arithmetic paths; other band settings include rare disagreements. This check does not establish arithmetic equivalence for arbitrary inputs.

Locality diagnostics use $R = \sqrt{2(1-\tau)}$, with cosine thresholds $\tau = 0.35/0.90/0.75/0.75$ for Sphere/Mixture/GloVe/COCO, respectively. Near pairs satisfy $d \le R$ and far pairs $d > R$, giving 8/9/4/6 near pairs. For each near pair, its actual distance supplies the margin threshold $2d(x, y)$ and the stronger endpoint supplies its stability bound. Far-pair bounds use the two endpoint coverage probabilities at radius $R/2$. These pair-specific bounds are integrated under the uniform 16-pair distribution, using Equation (45) for collision precision. This stated pair mixture is not a corpus-wide prior. Posterior error bars use the encoder-level delta method for a ratio of joint-probability estimates.

C.2 Semantic-pair screening

Data and representation. MRPC uses the GLUE train/development/test splits of 3,668/408/1,725 pairs. PAWS-Wiki Labeled (Final) uses 49,401/8,000/8,000 pairs. QQP reserves the 40,430 public-validation pairs for held-out evaluation; a fixed 90/10 split of public training gives 327,461 training and 36,385 development pairs. All retain their natural label prevalence. The shared frozen Sentence Transformers `paraphrase-MiniLM-L6-v2` encoder (Reimers & Gurevych, 2019) produces normalized 384-dimensional vectors with maximum sequence length 128. No encoder is fine-tuned. Its documented training sources include Quora duplicates; instance-level overlap with QQP evaluation has not been audited. Development and held-out QQP also share some source strings, so these are held-out *pairs*, not a text-disjoint generalization test.

Semantic screening and parameter selection. Strict training references contain 6,768 MRPC vectors and 40,223 PAWS vectors after excluding held-out texts and bit-identical embeddings. QQP uses the same text-disjoint, uniformly sampled 32,768-reference set; different texts with equal embeddings may remain. The cosine threshold maximizes development F1 on 401 values in $[-1, 1]$, with the largest-threshold tie break, yielding 0.640/0.750/0.745 for MRPC/PAWS/QQP.

We fix $T = 100$ and search $\psi \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$. The signature always has exactly T nearest-reference indices. We use $b \in \{1, 2, 4, 5, 10, 20, 25, 50\}$, with no discarded coordinates, giving 80 configurations per fit. A coordinate-count control directly bands 100 independent Gaussian-projection signs using the same eight widths. The primary hyperplane comparison instead matches theoretical signature capacity: after freezing SMinHash's development-selected ψ, it uses $q = T\lceil \log_2 \psi \rceil$ independent signs. It searches every positive divisor of q no greater than 50, using all q coordinates. This matches theoretical bits, not packed memory or runtime.

Both retention targets, 0.95 and 0.99, require feasibility in every fitted encoder on development data. A fixed sample of 65,536 probe cross pairs screens candidate selectivity, while the complete labeled development pairs determine retention. For SMinHash and the 100-sign control, at most 12 configurations per family are confirmed on the complete probe, first preserving the best feasible point at each reference-set size and then filling by increasing mean sampled candidate fraction. The capacity-matched hyperplane control confirms every allowed band width on the full probe, without shortlisting. Selection minimizes mean complete-probe candidate fraction subject to this constraint, separately for each target, before held-out evaluation. The targets share signatures; they are not independent fits. The benchmark splits had previously been inspected, so this parameter study is not an unseen-data confirmation. Figure 3 compares the two frozen targets; selected configurations accompany the reproducibility artifacts.

Each cost probe is the full cross product of at most 2,000 selected records from the corresponding split; MRPC uses 408 development records and 1,725 held-out records. The probe is distinct from the labeled-pair population and does not constitute a full-corpus join. We report sample SD over

three fitted hashes. Conditional official-pair bootstrap intervals are retained for F1; shared sentences can violate independent-pair sampling. These results concern candidate selectivity and semantic-rule retention, not end-to-end corpus deduplication time or statistical equivalence of classifiers.

C.3 VECTOR RETRIEVAL, GRAPHS, AND RANGE JOINS

Data and splits. We use GloVe25, GloVe100, COCO512, SIFT128, Deep96, GIST960, NYTimes256, and LastFM65 in the ANN-Benchmarks format (Aumüller et al., 2020). COCO uses the image-to-image CLIP ViT-B/16 features (Radford et al., 2021) from the random-split v0.1.3 release,[2] with 113,287 source database images and 10,000 image queries; the remaining collections use the ANN-Benchmarks distributions. The two GloVe dimensions share a corpus family. LastFM pairs user queries with item vectors, retaining the publisher's extra coordinate for angular retrieval. SIFT and GIST keep their native Euclidean metric. Angular vectors are explicitly normalized. The NYTimes view removes 239 database and nine query zero vectors, leaving 289,761 database items and 9,991 queries. We recompute exact neighbors after sampling and preprocessing.

The common comparison uses 100,000 database vectors, 200 validation queries, and 1,000 test queries (800 for GIST). Fixed permutations define nested sizes. These benchmark splits were previously inspected. GloVe100 and LastFM have separately fixed disjoint validation/test query subsets. References are the first 16,384 database items; IVF training uses a fixed sample of 65,536 items. Queries never enter these fitting samples. Scaling freezes the 100k choices, references, training items, fitted encoders and query IDs as Deep grows to one million and 9.99 million items; exact top-10 truth is recomputed at each size.

Selection for banded retrieval. The $T = 100$ grid above is also used for retrieval. For each reference encoder and seed, we evaluate the exact band gate on the 200 validation queries' true ten nearest neighbors. A fixed sample of 8,192 random query/database pairs estimates background candidate fraction. These screening estimates are distinct from complete-index candidate counts and runtime measurements. For each hashing seed and target in $\{0.90, 0.95, 0.99\}$, we shortlist the two feasible configurations with smallest probe fraction and one with the smallest encoding-work proxy. The encoding-work proxy is $T\psi d$. Infeasible screens retain a highest-retention diagnostic only. Deduplicated finalists undergo complete validation search, including all postings and exact reranking. This bounded shortlist is not an exhaustive runtime optimum.

Within confirmed points, minimum query time and minimum candidate fraction are selected separately at each target and frozen before test. Table 7 uses the minimum-time 0.95 selections with the shared optimized executor.

Baseline coverage. Retrieval includes HNSW (Malkov & Yashunin, 2020), FAISS IVF-Flat, IVF-PQ, and Flat (Douze et al., 2024; Johnson et al., 2021; Jégou et al., 2011), and geometric LSH. HNSW searches $M \in \{16, 32\}$ and `efSearch`$\in \{20, 50, 100, 200, 400, 800\}$ with `efConstruction`=200. IVF searches 256/1,024 lists and 1/4/16/64/256 probes. PQ uses 8-bit subquantizers and 16 code bytes (48 for GIST), without full-vector reranking. Search parameters are validation-selected; one fit supports all search settings of the same build configuration. Geometric LSH uses random hyperplanes for angular vectors (Charikar, 2002) and Gaussian p-stable hashes for Euclidean vectors (Datar et al., 2004). Both use 16 bands: hyperplane band widths are $\{4, 8, 12, 16\}$; p-stable widths are $\{2, 4, 8\}$ with bucket-scale multipliers $\{0.25, 0.5, 1\}$ of the median distance over 2,048 sampled reference pairs. Infeasible and unavailable selections remain explicit.

Graph and join selection. The task-specific selections, disjoint validation vertices and complete output criteria are described in Appendix C.3.4. Graph recall is evaluated on 1,024 held-out vertices; join recall uses all exact threshold edges. Complete execution time covers all 100,000 vertices.

Timing, uncertainty, and limits of the original banded studies. GPU methods use an RTX 5880 Ada; HNSW uses eight CPU threads on a dual Xeon Platinum 8352V host. Inputs are resident before timing. Hash encoding uses FP32 with TF32 disabled, while truth and hash-candidate verification use FP64. FAISS and HNSW retain native score precision, so strict-ID recall can differ from one

[2] `https://github.com/fabiocarrara/str-encoders/releases/tag/v0.1.3`

even for exhaustive FP32 search near ties. Construction includes fitting and insertion. Every timed hash query includes encoding, lookup, candidate union, and verification. Validation uses the median of two repetitions, and test uses the median of three; 32 individual-query latencies are also recorded. Graph and join totals include construction. Loading, embedding inference, initial transfers and oracle creation are excluded. The GPU is shared: other registered compute processes were observed during graph/join runs, with activity unresolved. Timings therefore include shared-environment variation and do not establish isolated-device throughput.

Results aggregate three fitted seeds with sample SD. Repeated Flat fits measure timing variation, not randomized sketch uncertainty. Index size includes reference coordinates, stored direct signatures, and postings; retained size is distinct from peak construction memory. Work exceeding 10^9 posting occurrences is censored before expansion, with no bucket truncation. Hash search has a 240-second deadline; exact range has a 600-second deadline and a 20-million-edge output limit. These method-dependent limits are not a common-budget speed comparison. The environment includes PyTorch 2.7.0, FAISS 1.12.0, HNSWlib 0.8.0, and CUDA 12.6.

C.3.1 Modern text and multimodal embeddings

DBpedia1536 uses Qdrant's published `text-embedding-3-large` vectors from entity titles and descriptions (Qdrant, 2024; OpenAI, 2024). COYO2048 uses Qdrant's published Qwen3-VL-Embedding-2B vectors from joint image–caption inputs (Qdrant, 2026; Li et al., 2026). These vector-neighbor tasks use neither BEIR relevance labels nor the COYO release's ImageNet query set.

Publisher revisions, lexicographic shard prefixes, and sampling IDs are fixed before measurement. Pools of 115,386 and 120,000 records, respectively, each supply 100,000 database items, 200 validation queries, and 1,000 test queries with disjoint source IDs. Both releases are losslessly representable in FP32 before normalization; all vectors are normalized. Exact FP64 top-ten neighbors are recomputed against each sampled database. References use the first 16,384 database rows; IVF training uses 65,536 fixed rows. Both collections follow the banded retrieval grids, screening, selection and timing protocol in Appendix C, with full-workload warmup for every family, two measured validation passes and three test passes. Across-fit means and sample SD use within-fit medians.

Held-out records come from these shard pools, not the publishers' full-corpus benchmarks. Embeddings are not regenerated; source revisions, checksums and split IDs accompany the code.

C.3.2 Sensitivity to geometric signature capacity

The primary geometric grid uses 16 bands. A post-hoc control expands band counts on both modern datasets using the same test queries, so it provides no independent confirmatory evidence.

For each dataset, seed, and validation target in $\{0.90, 0.95, 0.99\}$, the original validation-selected SMinHash configuration is retained. Its nominal signature capacity caps the search at $C = 100\lceil \log_2 \psi \rceil$ bits. Hyperplane LSH searches band widths $b \in \{4, 6, 8, 10, 12, 14, 16, 20, 24, 32\}$ and band counts $L \in \{16, 32, 64, 128, \lfloor C/b \rfloor\}$, retaining $L \geq 1$ and $bL \leq C$. Every binary coordinate belongs to a complete band. This nominal capacity cap controls neither index memory, computation nor tuning effort, and establishes no optimal geometric frontier.

Each setting receives a complete 200-query warmup and three validation batches; selection minimizes median query time among target-feasible settings. Both families are rebuilt and timed with a full 1,000-query warmup and three rotating test batches; identical configurations share a fit across targets. Test outcomes never change selections; infeasible and resource-censored results remain included. Unchanged SMinHash configurations reproduce their original neighbor IDs.

C.3.3 Product duplicate-candidate blocking

WDC Block's medium setting supplies product offers and labeled matching pairs, with 5,000 records in table A and 200,000 in table B (Brinkmann et al., 2023; 2024). Every A record queries B; candidate fraction is the number of distinct proposed record pairs divided by 10^9. Identical serialized texts share embedding computation but remain distinct records in the workload.

All methods use the same frozen, normalized 384-dimensional `paraphrase-MiniLM-L6-v2` representation (Reimers & Gurevych, 2019). Fields are serialized in title, brand, description, price, and currency order. The encoder's 128-token limit truncates 61,825 of the 200,616 unique texts (30.82%). Encoders are not fine-tuned. Results concern this common, length-limited representation.

Removing development records and serialized texts overlapping the official test set leaves 7,142 development pairs, including 2,751 positives, spanning 3,140 distinct A queries. The test set contains 4,500 labeled pairs, including 500 positives. Development/test records, texts, and pairs are disjoint. We do not infer additional matching labels from incomplete cluster identifiers. Recall measures the fraction of official test positives proposed; candidate volume uses the complete A-by-B workload.

References are 16,384 unlabeled background B texts outside official development/test records, texts, recorded clusters, and positive-pair components. This database-dependent sampling cannot certify complete entity-level separation from the encoder's training corpus or all background records.

For each of three seeds, development selection minimizes candidate volume subject to positive-pair recall at least 0.95 or 0.99. SMinHash screens all 80 direct configurations; the geometric control screens 51 sign-and-band configurations. A fixed random-pair probe estimates background candidate fraction before complete development-query evaluation of shortlisted configurations. ANN blockers jointly select index settings and returned-neighbor count. The final common return-count grid is

$$R \in \{10, 20, 50, 100, 200, 500, 1000, 1250, 1500, 2000, 3000, 4000, 5000, 10^4, 2{\cdot}10^4\};$$

the grid and final configurations are fixed before corresponding test evaluation.

Hash and Flat blocking use the RTX 5880 Ada GPU; HNSW and the full-return-count IVF adapter use eight CPU threads. Blocking time covers native candidate generation, including query hashing, separately from index construction. Complete-workload warmup precedes three measured passes; their medians yield the mean and sample SD across three fits. Optional common FP64 candidate scoring is reported separately without redefining official matches. Inputs reside on each index's device; timings exclude candidate transfers and audit overhead. Flat has zero reported build cost because it constructs no separate index. Infeasible selections and resource failures remain included; unmet targets characterize the declared search space, not algorithm-family impossibility.

C.3.4 Task-specific validation for graphs and joins

Graphs and threshold joins use task-specific validation; complete range comparisons below retain these definitions and selected reference and band parameters.

Each 100,000-vector database supplies 1,024 fixed validation vertices and 1,024 disjoint test-audit vertices; the unlabelled database remains available for reference sampling. Graphs query every database vertex, exclude self matches, and retain ten neighbors; held-out Recall@10 measures accuracy. Threshold joins return every generated directed non-self pair passing exact metric verification. Edge recall and precision use the complete exact threshold-edge set whenever that reference finishes.

Before validation, 100,000 fixed non-self pairs calibrate the threshold to a prospectively chosen pair-density target of 10^{-4}; realized full-join density is not guaranteed. We screen all 80 combinations of $T = 100$, $\psi \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$ and eight complete band widths. Complete index queries confirm at most six feasible settings per seed and task, ordered by estimated candidate fraction. Geometric controls use all four angular or nine Euclidean settings; graph ANN controls use twelve HNSW and ten IVF-Flat settings, plus Flat. All families share task validation truth. Subject to recall at least 0.95, hash methods minimize confirmed candidate fraction and graph ANN methods minimize query time. Selections freeze before held-out task truth is computed.

Both hash families reuse signatures from the full database build. Each complete validation workload is warmed once and timed three times; each complete test workload is warmed and timed once per independently fitted seed. Reported construction-plus-query time excludes separately recorded warmup. HNSW uses eight CPU threads; other methods use the RTX 5880 Ada GPU, with inputs resident on the index device. Full warmups and measured calls share per-call limits: 240 seconds for methods, 600 seconds for the exact join, 10^9 posting occurrences and 2×10^7 output edges. Cooperative checks censor budget-exceeding calls without truncation; failed warmup ends the measurement attempt.

All planned seed outcomes remain included; incomplete means report coverage and cannot become bold winners. If a full join or its reference is censored, a separately identified $1{,}024 \times N$ held-out range search may be measured, never substituted for a complete join. The separate retrieval-to-task transfer control retains its original tenth-neighbor threshold, graph audit vertices, encoding policy and timing protocol. Differences across studies cannot be attributed solely to task-specific parameter selection.

C.3.5 COMPLETE GPU-ASSISTED RANGE JOINS

SMinHash, IVF-Flat range search, and exhaustive verification share COCO, Deep, SIFT and NYTimes 100k databases, thresholds, 1,024 validation vertices and complete FP64 truth sets. Each approximate method has three independent fits, with parameters selected on validation vertices before complete-join evaluation. This complete-cost comparison reuses the task-development cohort.

Candidate membership. FAISS 1.12.0 trains and searches IVF-Flat on the GPU. Its range helper retrieves a capacity-limited nearest-neighbor list; queries whose last returned score passes the threshold receive uncapped CPU range search on a clone of the GPU-trained index with the same probe count. Capacity controls work without limiting final neighbors. Mandatory CPU overflow is included in cost; the search is GPU-assisted. The native FP32 gate uses squared Euclidean distance for SIFT and inner product on pre-normalized angular vectors, without expanding its strict threshold. The common inclusive FP64 scorer verifies every returned non-self pair. Native rounding or coarse routing can omit true edges; accuracy compares all verified directed non-self edges against full FP64 truth.

Validation grid. IVF uses 256 or 1,024 lists with every power-of-two probe count from one through all lists. Training uses $\max(16{,}384, 40n_{\text{list}})$ or 65,536 database vectors, with 10 or 20 clustering iterations; the smaller set is a prefix of the same fixed draw. Helper capacities $\{128, 512\}$ and query batches $\{256, 1024\}$ give 320 configurations per fit, reusing fitted indexes without retraining. SMinHash retains its selected $T = 100$ reference and band parameters, varying only execution batches over $\{32, 128, 512, 1024\}$. Workspace limits may split batches without truncating candidates. Construction encodes every database vertex; join queries reuse those signatures.

Workload objectives. Subject to validation edge recall at least 0.95, the query objective minimizes median validation query time V; the one-join objective minimizes $B + (N/1024)V$, with measured construction time B and $N = 100{,}000$. The projection is a selection proxy without claiming exact linear scaling. Lexicographic parameter order breaks ties. Both objectives are fixed before full evaluation; identical selections share one measurement. SMinHash's selections coincide because construction is shared across execution batches. Complete-join results never revise selections.

Timing and aggregation. Pipelines run serially on one RTX 5880 Ada GPU on the shared host, with eight Intel Xeon Platinum 8352V CPU threads for overflow. Every validation and complete workload receives one full warmup and three timed passes. Tables summarize within-fit medians by across-fit means and sample SD; exhaustive verification uses one configuration and three passes. GPU IVF construction includes initialization, training, insertion and the CPU index clone. Query cost includes gathering, host/device transfers, CPU overflow, FP64 verification and host output; it excludes initial bulk database staging, post-query sorting, serialization and accuracy calculation. Query limits are 240 seconds, or 600 for exhaustive verification, and 20 million output edges; failures are censored without truncation. Totals cover post-selection construction and one complete join, excluding both families' validation-search cost; achieved recalls can differ. Table 12 reports the one-join objective, including CPU overflow and full-join recall-target failures; Appendix D.8 also summarizes the query-only objective. Execution-batch changes exactly reproduce original SMinHash edge IDs and candidate counts in every fit.

D TASK COMPARISONS AND PARAMETER SENSITIVITY

The comparisons below retain all dataset and baseline families. Unless stated otherwise, results are means and sample standard deviations over three fits. Infeasible, unavailable and resource-censored

Table 4: **Candidate selectivity depends on the embedding dataset.** Mean$_{\pm SD}$ over three fits on 1,000 new queries per 100k database. Retention is the percentage of FP64 top-10 neighbors admitted; Candidates is the percentage of the database admitted. Superscripts count fits below 95% retention. Bold marks the higher retention mean; candidate minima require 95% retention in all three fits. Displayed ties are included. Parameters were fixed before these queries were evaluated.

	SMinHash		HP1000	
Dataset	Retention (%)	Candidates (%)	Retention (%)	Candidates (%)
Deep96	$\mathbf{95.72}^{0}_{\pm 0.70}$	$0.974_{\pm 0.021}$	$95.70^{0}_{\pm 0.10}$	$\mathbf{0.925}_{\pm 0.046}$
COCO512	$94.45^{3}_{\pm 0.47}$	$0.462_{\pm 0.027}$	$\mathbf{95.83}^{1}_{\pm 0.89}$	$1.602_{\pm 0.249}$
GloVe25	$\mathbf{95.52}^{0}_{\pm 0.15}$	$\mathbf{0.922}_{\pm 0.003}$	$94.15^{3}_{\pm 0.47}$	$0.442_{\pm 0.114}$
GloVe100	$94.48^{3}_{\pm 0.30}$	$5.618_{\pm 0.182}$	$\mathbf{94.57}^{2}_{\pm 0.92}$	$1.042_{\pm 0.131}$
LastFM65	$\mathbf{95.82}^{0}_{\pm 0.52}$	$\mathbf{0.332}_{\pm 0.027}$	$95.26^{1}_{\pm 0.74}$	$45.298_{\pm 3.606}$
NYTimes256	$95.37^{0}_{\pm 0.23}$	$11.578_{\pm 1.032}$	$\mathbf{95.66}^{0}_{\pm 0.31}$	$\mathbf{0.801}_{\pm 0.076}$
DBpedia1536	$93.68^{3}_{\pm 1.04}$	$0.758_{\pm 0.114}$	$\mathbf{94.61}^{2}_{\pm 0.58}$	$0.196_{\pm 0.018}$
COYO2048	$94.51^{3}_{\pm 0.27}$	$0.659_{\pm 0.001}$	$\mathbf{94.59}^{3}_{\pm 0.25}$	$0.250_{\pm 0.027}$

results retain their coverage and are not interpreted as zero cost or zero recall. Protocols are specified in Appendix C.

D.1 New-query confirmation and retention-target sensitivity

Selection and query separation. The primary confirmation in Table 4 fixes all choices from the original 200 validation queries per dataset. For each of three seeds and each $\psi \in \{2, 4, \ldots, 1024\}$, we choose the largest integer m whose count gate retains at least 95% of the 2,000 validation top-10 neighbors. We then select the feasible resolution with the smallest complete candidate count, breaking ties toward smaller ψ. Hyperplane1000 uses the same threshold-selection criterion. Both families share the normalized vectors and numerical truth, but their encoding work, model storage, and tuning budgets differ.

Each dataset contributes 1,000 additional queries selected before evaluation and excluded from all earlier development and test query IDs. The database and reference collection remain unchanged. Each fitted encoder evaluates this same query sample; three fits do not create 3,000 independent queries. Original database signatures are reused. Queries are encoded in requests of 32, with a final request of eight, to preserve the prescribed floating-point execution shape. Candidate counts cover every query–database pair, without candidate truncation. The reported retention measures membership of numerical top-10 neighbors in the candidate set, without reranking or timing.

Numerical truth. Top-10 truth uses the canonical FP64 score and a deterministic database-ID tie break. Query and database tiles are fixed throughout the comparison, and stable merging preserves the lexicographic score/ID order. This is a specified floating-point reference calculation, not exact real arithmetic. Independent numerical checks and executable settings accompany the code.

Primary count-gate comparison. At the primary target, SMinHash has a lower candidate fraction on LastFM. On GloVe25, only SMinHash meets the mean retention target; its candidate fraction is higher. All its fits meet the target on those two datasets. Among families meeting the mean target, Hyperplane1000 is more selective on Deep, COCO and NYTimes; neither meets it in mean on GloVe100, DBpedia or COYO. Each family misses the target in 12 of 24 fits. The LastFM centering control and its representation-specific explanation appear in Appendix E.3. These candidate-quality results are distinct from returned retrieval in Table 18, which preserves the complete four-family timing and recall comparison on these fixed queries.

Predeclared secondary targets. The 90% and 99% targets were specified before the primary query evaluation. Their parameters come from the same original validation grid, with no selection on the newly evaluated queries. Table 5 reports both targets on that same query sample. Model caches shared with the primary study are reused; changing a threshold is not another encoder fit or independent trial.

Table 5: Sensitivity to the predeclared retention target. The same 1,000 queries per dataset as Table 4 evaluate parameters selected on the original validation queries. Values are percentages, mean$_{\pm SD}$ across three hash seeds; superscripts count fits below the stated target. Bold marks the higher retention mean; candidate minima require unrounded mean retention to meet the target. Displayed ties are included. All 96 target results are included. Reused queries and model caches are not independent additional trials.

		SMinHash		HP1000	
Dataset	Target	Retention (%)	Candidates (%)	Retention (%)	Candidates (%)
Deep96	90%	$\mathbf{91.67}^{0}_{\pm 0.94}$	$\mathbf{0.526}_{\pm 0.048}$	$89.80^{2}_{\pm 0.50}$	$0.560_{\pm 0.031}$
	99%	$99.16^{1}_{\pm 0.17}$	$2.738_{\pm 0.244}$	$\mathbf{99.42}^{0}_{\pm 0.08}$	$\mathbf{1.877}_{\pm 0.071}$
COCO512	90%	$88.67^{3}_{\pm 0.76}$	$0.277_{\pm 0.013}$	$\mathbf{90.60}^{0}_{\pm 0.62}$	$\mathbf{0.876}_{\pm 0.086}$
	99%	$99.00^{1}_{\pm 0.09}$	$1.138_{\pm 0.004}$	$\mathbf{99.04}^{2}_{\pm 0.28}$	$\mathbf{4.184}_{\pm 0.950}$
GloVe25	90%	$\mathbf{90.28}^{0}_{\pm 0.24}$	$\mathbf{0.490}_{\pm 0.041}$	$89.03^{3}_{\pm 0.73}$	$0.295_{\pm 0.081}$
	99%	$\mathbf{99.04}^{1}_{\pm 0.14}$	$\mathbf{2.266}_{\pm 0.007}$	$98.40^{3}_{\pm 0.31}$	$0.815_{\pm 0.245}$
GloVe100	90%	$\mathbf{91.13}^{0}_{\pm 0.24}$	$\mathbf{3.619}_{\pm 0.119}$	$88.99^{2}_{\pm 1.30}$	$0.608_{\pm 0.057}$
	99%	$\mathbf{98.52}^{3}_{\pm 0.26}$	$13.834_{\pm 1.421}$	$98.48^{3}_{\pm 0.22}$	$2.184_{\pm 0.126}$
LastFM65	90%	$\mathbf{90.89}^{0}_{\pm 0.97}$	$\mathbf{0.207}_{\pm 0.013}$	$90.60^{1}_{\pm 1.27}$	$30.721_{\pm 3.204}$
	99%	$\mathbf{99.16}^{0}_{\pm 0.14}$	$\mathbf{0.857}_{\pm 0.113}$	$99.00^{1}_{\pm 0.53}$	$70.586_{\pm 5.218}$
NYTimes256	90%	$\mathbf{91.46}^{0}_{\pm 0.34}$	$5.732_{\pm 0.639}$	$91.41^{0}_{\pm 0.43}$	$\mathbf{0.367}_{\pm 0.044}$
	99%	$\mathbf{99.34}^{0}_{\pm 0.25}$	$\mathbf{36.527}_{\pm 4.168}$	$98.85^{2}_{\pm 0.36}$	$2.501_{\pm 0.699}$
DBpedia1536	90%	$88.49^{3}_{\pm 0.03}$	$0.385_{\pm 0.001}$	$\mathbf{89.54}^{3}_{\pm 0.53}$	$0.107_{\pm 0.007}$
	99%	$98.32^{3}_{\pm 0.17}$	$2.400_{\pm 0.066}$	$\mathbf{98.98}^{2}_{\pm 0.18}$	$0.614_{\pm 0.028}$
COYO2048	90%	$88.29^{3}_{\pm 0.93}$	$0.348_{\pm 0.033}$	$\mathbf{89.69}^{2}_{\pm 0.87}$	$0.158_{\pm 0.018}$
	99%	$98.82^{2}_{\pm 0.24}$	$1.996_{\pm 0.083}$	$\mathbf{99.00}^{2}_{\pm 0.18}$	$0.680_{\pm 0.096}$

LastFM's candidate reduction persists at both targets: the ratio of Hyperplane1000 to SMinHash mean candidate fractions is 148.13 at 90% and 82.33 at 99%, with both families meeting each target in aggregate. All three SMinHash fits meet both targets on LastFM. COCO has 3.68 times fewer candidates with SMinHash at the 99% selection target, but its mean retention remains just below 99%, so it is ineligible under the strict target despite appearing equal at lower display precision. These results retain the distinction between measured selectivity and target attainment. Across all 24 fits per family, the 90% target is missed by 9 SMinHash fits and 13 hyperplane fits; at 99%, the counts are 11 and 15. These secondary observations supplement the primary 95% result rather than replacing its failures.

D.2 Reference resolution and signature capacity

Figure 3 compares direct banding of reference indices with $T\lceil \log_2 \psi \rceil$ independent hyperplane signs, matching nominal bit capacity at development targets 0.95 and 0.99. Encoding cost, physical memory and achieved test retention can differ. On PAWS at target 0.95, candidate fractions are close (0.137% versus 0.141%), but hyperplanes retain more cosine-positive pairs (98.20% versus 96.27%). Thus SMinHash does not uniformly dominate this control. A 100-sign control is less selective than the capacity-matched control on all three datasets at both targets; its mean retention is 98.341–100.000%. The smaller code does not isolate reference selection at equal nominal capacity.

D.3 Product blocking and downstream candidate cost

Table 6 gives the 95% development target; the 99% sensitivity result is summarized below. All three fits complete for every family at both targets. Recall covers official positive pairs; the full 10^9 record-pair workload supplies candidate volume, not negative labels.

At the stricter target, SMinHash reaches 99.13% test recall with 41.36% candidates, versus 99.60% and 67.27% for hyperplanes. Top-R vector search is more selective: 10% candidates at 99.60% recall. A shared development constraint therefore does not imply equal test recovery.

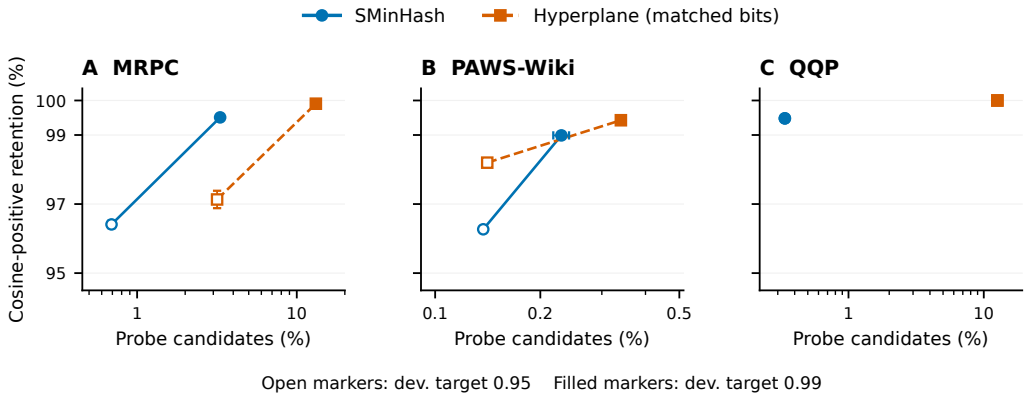

Figure 3: **Reference-based signatures support selective semantic screening.** Held-out cosine-positive retention versus candidate fraction on the separate cross-product probe. Points summarize three fits, with one sample SD; segments connect the configurations selected at development targets 0.95 and 0.99. They are frozen operating points, not frontiers optimized on test outcomes. Both families apply banding directly to their signatures. Different achieved retention must be considered alongside candidate cost. Hyperplane signatures match theoretical bit capacity; actual packed memory and runtime are not matched.

Table 6: WDC blocking at a 95% development recall target, which does not imply equal test recall. Candidates span $5{,}000 \times 200{,}000 = 10^9$ record pairs; recall covers 500 official test positives. Query times include query hashing and are medians of three complete passes after full-workload warmup. Entries give mean$_{\pm \mathrm{SD}}$ over three seeds. GPU: RTX 5880 Ada; CPU: eight threads. Text encoding and optional scoring are excluded. Bold marks best complete means, with time minima compared within CPU or GPU rows and ties included. Cost minima require mean recall at least 95%; achieved recalls can differ.

Method	Recall (%) ↑	Candidates (%) ↓	Build (s) ↓	Query (s) ↓
SMinHash (GPU)	$96.60_{\pm 0.35}$	$6.225_{\pm 0.046}$	$1.493_{\pm 0.272}$	$\mathbf{0.169}_{\pm 0.039}$
Hyperplane (GPU)	$\mathbf{97.13}_{\pm 0.70}$	$31.423_{\pm 0.061}$	$0.088_{\pm 0.012}$	$0.178_{\pm 0.044}$
Flat (GPU)	$96.60_{\pm 0.00}$	$\mathbf{0.625}_{\pm 0.000}$	$\mathbf{0.000}_{\pm 0.000}$	$0.183_{\pm 0.038}$
HNSW (CPU)	$96.60_{\pm 0.00}$	$\mathbf{0.625}_{\pm 0.000}$	$28.887_{\pm 1.185}$	$\mathbf{6.887}_{\pm 0.283}$
IVF-Flat (CPU)	$96.60_{\pm 0.00}$	$\mathbf{0.625}_{\pm 0.000}$	$\mathbf{21.350}_{\pm 1.408}$	$39.671_{\pm 0.734}$

An optional scoring control applies the same FP64 cosine-distance kernel to every candidate. At the 95% target, this costs 0.525 seconds for SMinHash, 2.801 for hyperplanes and 0.063 for Flat. This optional pass measures vector-scoring work, not entity classification. Its sum with construction and blocking excludes text encoding, candidate transfers and audit overhead (Appendix C.3.3).

D.4 Retrieval across baseline families

Table 7 compares all six families at the 0.95 validation target. Both LSH families use the same optimized executor; HNSW uses eight CPU threads and the others use the RTX 5880 Ada GPU. Appendix F.1 distinguishes query costs from construction costs and reports the resource-reuse control.

D.5 Text and multimodal retrieval

DBpedia1536 and COYO2048 measure vector-neighbor recovery on disjoint records from fixed text and joint image–text embedding pools (Appendix C.3.1). Tables 8 and 9 retain every family at the 0.95 validation target, including infeasible selections.

Recovery and cost across targets. The same frozen selections also evaluate the 0.90 and 0.99 targets; no test outcome changes a selection. At target 0.95, mean SMinHash Recall@10 is 98.81%

Table 7: **Construction and query cost give different rankings.** Both LSH methods share the optimized executor (Appendix F.1); all indexes are rebuilt and retimed. Total adds construction and one complete query batch. Values are mean$_{\pm \mathrm{SD}}$ over three fits; NF denotes infeasible validation selection and uninstrumented ANN candidates.

Dataset	Method	Recall	Cand. (%)	Build (s)	Query (ms)	Total (s)
GloVe25	SMinHash	$0.958_{\pm 0.001}$	$\mathbf{2.99}_{\pm 0.03}$	$0.03_{\pm 0.00}$	$14.7_{\pm 0.8}$	$\mathbf{0.043}_{\pm 0.004}$
	Geo. LSH	$0.981_{\pm 0.032}$	$59.84_{\pm 33.99}$	$\mathbf{0.00}_{\pm 0.00}$	$72.1_{\pm 29.2}$	$0.075_{\pm 0.028}$
	Flat (GPU)	$\mathbf{1.000}_{\pm 0.000}$	–	$0.18_{\pm 0.04}$	$2.2_{\pm 0.0}$	$0.178_{\pm 0.035}$
	HNSW (CPU)	$0.970_{\pm 0.000}$	–	$4.29_{\pm 0.20}$	$18.8_{\pm 0.6}$	$4.305_{\pm 0.200}$
	IVF-Flat	$0.959_{\pm 0.015}$	–	$0.25_{\pm 0.01}$	$\mathbf{1.7}_{\pm 0.1}$	$0.248_{\pm 0.010}$
	IVF-PQ	NF	NF	NF	NF	NF
COCO512	SMinHash	$0.968_{\pm 0.001}$	$\mathbf{1.10}_{\pm 0.00}$	$0.34_{\pm 0.00}$	$16.0_{\pm 0.3}$	$0.361_{\pm 0.004}$
	Geo. LSH	$0.958_{\pm 0.006}$	$47.79_{\pm 2.64}$	$\mathbf{0.01}_{\pm 0.00}$	$307.3_{\pm 14.5}$	$0.313_{\pm 0.014}$
	Flat (GPU)	$\mathbf{1.000}_{\pm 0.000}$	–	$0.14_{\pm 0.00}$	$5.0_{\pm 0.0}$	$\mathbf{0.148}_{\pm 0.004}$
	HNSW (CPU)	$0.961_{\pm 0.006}$	–	$10.43_{\pm 0.34}$	$27.6_{\pm 2.5}$	$10.463_{\pm 0.345}$
	IVF-Flat	$0.977_{\pm 0.002}$	–	$0.96_{\pm 0.02}$	$\mathbf{3.9}_{\pm 0.0}$	$0.967_{\pm 0.015}$
	IVF-PQ	NF	NF	NF	NF	NF
SIFT128	SMinHash	$0.953_{\pm 0.002}$	$\mathbf{1.98}_{\pm 0.01}$	$0.06_{\pm 0.00}$	$13.4_{\pm 0.1}$	$\mathbf{0.071}_{\pm 0.000}$
	Geo. LSH	$0.954_{\pm 0.002}$	$33.59_{\pm 1.87}$	$\mathbf{0.01}_{\pm 0.00}$	$84.9_{\pm 4.9}$	$0.091_{\pm 0.005}$
	Flat (GPU)	$\mathbf{1.000}_{\pm 0.000}$	–	$0.16_{\pm 0.01}$	$\mathbf{2.3}_{\pm 0.2}$	$0.159_{\pm 0.009}$
	HNSW (CPU)	$0.982_{\pm 0.007}$	–	$5.66_{\pm 0.45}$	$27.3_{\pm 3.1}$	$5.687_{\pm 0.451}$
	IVF-Flat	$0.988_{\pm 0.014}$	–	$0.43_{\pm 0.03}$	$2.6_{\pm 1.0}$	$0.433_{\pm 0.034}$
	IVF-PQ	NF	NF	NF	NF	NF
Deep96	SMinHash	$0.973_{\pm 0.019}$	$\mathbf{3.28}_{\pm 0.31}$	$0.10_{\pm 0.11}$	$20.7_{\pm 0.7}$	$\mathbf{0.118}_{\pm 0.114}$
	Geo. LSH	$0.996_{\pm 0.001}$	$69.65_{\pm 0.98}$	$\mathbf{0.00}_{\pm 0.00}$	$137.2_{\pm 1.5}$	$0.141_{\pm 0.002}$
	Flat (GPU)	$\mathbf{1.000}_{\pm 0.000}$	–	$0.16_{\pm 0.01}$	$2.3_{\pm 0.0}$	$0.158_{\pm 0.008}$
	HNSW (CPU)	$0.974_{\pm 0.000}$	–	$4.40_{\pm 0.58}$	$20.0_{\pm 0.4}$	$4.420_{\pm 0.577}$
	IVF-Flat	$0.982_{\pm 0.011}$	–	$0.33_{\pm 0.01}$	$\mathbf{2.0}_{\pm 0.1}$	$0.336_{\pm 0.008}$
	IVF-PQ	NF	NF	NF	NF	NF
GIST960	SMinHash	$0.987_{\pm 0.001}$	$\mathbf{11.64}_{\pm 0.04}$	$1.35_{\pm 0.03}$	$144.4_{\pm 2.6}$	$1.491_{\pm 0.028}$
	Geo. LSH	$0.988_{\pm 0.001}$	$88.41_{\pm 1.93}$	$\mathbf{0.01}_{\pm 0.00}$	$833.7_{\pm 19.3}$	$0.840_{\pm 0.019}$
	Flat (GPU)	$\mathbf{1.000}_{\pm 0.000}$	–	$0.17_{\pm 0.01}$	$\mathbf{6.6}_{\pm 0.2}$	$\mathbf{0.178}_{\pm 0.007}$
	HNSW (CPU)	$0.963_{\pm 0.000}$	–	$27.82_{\pm 1.76}$	$202.2_{\pm 4.7}$	$28.022_{\pm 1.756}$
	IVF-Flat	$0.958_{\pm 0.002}$	–	$1.72_{\pm 0.14}$	$26.5_{\pm 0.2}$	$1.747_{\pm 0.139}$
	IVF-PQ	NF	NF	NF	NF	NF
NYTimes256	SMinHash	$0.974_{\pm 0.001}$	$\mathbf{26.27}_{\pm 0.05}$	$\mathbf{0.09}_{\pm 0.01}$	$108.6_{\pm 0.7}$	$0.201_{\pm 0.006}$
	Geo. LSH	NF	NF	NF	NF	NF
	Flat (GPU)	$0.995_{\pm 0.000}$	–	$0.17_{\pm 0.02}$	$\mathbf{3.4}_{\pm 0.1}$	$\mathbf{0.176}_{\pm 0.017}$
	HNSW (CPU)	$0.966_{\pm 0.002}$	–	$21.61_{\pm 0.71}$	$487.4_{\pm 16.2}$	$22.097_{\pm 0.700}$
	IVF-Flat	$\mathbf{0.996}_{\pm 0.000}$	–	$0.63_{\pm 0.01}$	$22.1_{\pm 0.1}$	$0.656_{\pm 0.010}$
	IVF-PQ	NF	NF	NF	NF	NF
GloVe100	SMinHash	$0.981_{\pm 0.001}$	$\mathbf{17.73}_{\pm 0.10}$	$0.12_{\pm 0.00}$	$61.3_{\pm 1.9}$	$0.177_{\pm 0.002}$
	Geo. LSH	$0.975_{\pm 0.004}$	$73.92_{\pm 2.48}$	$\mathbf{0.00}_{\pm 0.00}$	$145.7_{\pm 5.5}$	$\mathbf{0.149}_{\pm 0.006}$
	Flat (GPU)	$\mathbf{1.000}_{\pm 0.000}$	–	$0.15_{\pm 0.01}$	$\mathbf{3.0}_{\pm 0.6}$	$0.151_{\pm 0.014}$
	HNSW (CPU)	$0.971_{\pm 0.000}$	–	$9.00_{\pm 0.35}$	$121.1_{\pm 6.6}$	$9.120_{\pm 0.357}$
	IVF-Flat	$0.971_{\pm 0.009}$	–	$0.42_{\pm 0.08}$	$4.1_{\pm 0.9}$	$0.422_{\pm 0.084}$
	IVF-PQ	NF	NF	NF	NF	NF
LastFM65	SMinHash	$0.996_{\pm 0.001}$	$\mathbf{1.43}_{\pm 0.02}$	$0.20_{\pm 0.00}$	$13.7_{\pm 0.9}$	$0.213_{\pm 0.002}$
	Geo. LSH	NF	NF	NF	NF	NF
	Flat (GPU)	$\mathbf{1.000}_{\pm 0.000}$	–	$\mathbf{0.14}_{\pm 0.00}$	$2.3_{\pm 0.2}$	$\mathbf{0.143}_{\pm 0.004}$
	HNSW (CPU)	$0.956_{\pm 0.001}$	–	$3.03_{\pm 0.16}$	$12.5_{\pm 2.4}$	$3.047_{\pm 0.166}$
	IVF-Flat	$0.977_{\pm 0.008}$	–	$0.28_{\pm 0.01}$	$\mathbf{2.1}_{\pm 0.2}$	$0.286_{\pm 0.008}$
	IVF-PQ	NF	NF	NF	NF	NF

Table 8: **Selective candidates can accompany different achieved recalls.** Cosine retrieval with minimum-query-time validation selection at target 0.95; means$_{\pm SD}$ over three fits. † marks any fit below the test target; NF means no feasible validation choice; — means candidate counts are uninstrumented. Geometric LSH uses hyperplanes; shading identifies $T = 100$ SMinHash. Flat uses FP32 scoring against FP64 strict-ID truth. Bold marks the best complete column mean, including displayed ties; candidate minima do not imply equal-quality comparisons.

Method	DBpedia1536		COYO2048	
	Recall (%)	Cand. (%)	Recall (%)	Cand. (%)
Flat (GPU)	$\mathbf{100.00}_{\pm 0.00}$	—	$\mathbf{100.00}_{\pm 0.00}$	—
HNSW (CPU)	$95.39_{\pm 0.10}$	—	$96.05^{\dagger}_{\pm 1.84}$	—
IVF-Flat	$94.52^{\dagger}_{\pm 0.04}$	—	$94.81^{\dagger}_{\pm 0.12}$	—
IVF-PQ	NF	NF	NF	NF
Geometric LSH	$96.27_{\pm 0.65}$	$72.702_{\pm 2.475}$	$97.39_{\pm 0.06}$	$71.904_{\pm .721}$
SMinHash	$98.81_{\pm 0.05}$	$\mathbf{4.043}_{\pm .011}$	$96.36^{\dagger}_{\pm 2.77}$	$\mathbf{3.022}_{\pm .925}$

Table 9: **Flat queries faster despite the hash methods' selectivity.** Costs for Table 8's frozen settings. Query time covers all 1,000 test queries, including hash encoding and original-vector verification. Values are mean$_{\pm SD}$ over three fits, using medians of three query repetitions. HNSW uses eight CPU threads; other methods use the RTX 5880 Ada GPU. NF denotes infeasible validation selection. Bold marks column minima, including displayed ties; achieved recalls differ.

Method	DBpedia1536		COYO2048	
	Build (s)	Query (ms)	Build (s)	Query (ms)
Flat (GPU)	$.159_{\pm .005}$	$\mathbf{15.3}_{\pm 0.4}$	$.180_{\pm .012}$	$\mathbf{19.4}_{\pm 0.8}$
HNSW (CPU)	$37.237_{\pm 2.284}$	$131.3_{\pm 6.8}$	$53.795_{\pm 6.266}$	$156.9_{\pm 36.9}$
IVF-Flat	$2.688_{\pm .031}$	$61.6_{\pm 1.4}$	$3.037_{\pm .047}$	$72.2_{\pm 1.0}$
IVF-PQ	NF	NF	NF	NF
Geometric LSH	$\mathbf{.009}_{\pm .001}$	$1909.1_{\pm 30.4}$	$\mathbf{.009}_{\pm .000}$	$1931.7_{\pm 26.0}$
SMinHash	$2.133_{\pm .009}$	$138.7_{\pm 0.8}$	$1.111_{\pm 1.446}$	$105.4_{\pm 37.6}$

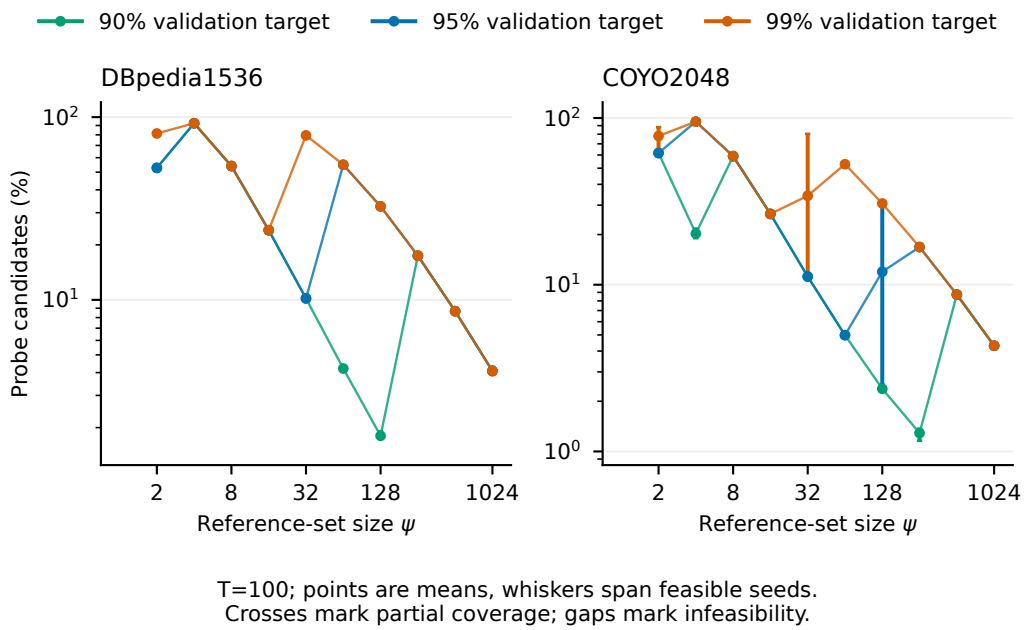

Figure 4: **Reference-set size controls candidate selectivity in validation-only screening.** At $T = 100$, each seed and ψ uses the lowest probe candidate fraction among band settings meeting the stated true-neighbor retention target. The probe contains 8,192 fixed random query/database pairs. Points show means and whiskers span feasible seeds; crosses mark partial coverage and gaps denote infeasibility. These are sampled candidate-gate estimates, not held-out retrieval measurements or runtime improvements.

on DBpedia and 96.36% on COYO, but two COYO fits miss the target. IVF-Flat misses it in all three fits on both datasets; HNSW misses once on COYO. At target 0.99 on COYO, SMinHash achieves 99.56% recall with 4.08% candidates and takes 153.8 ms per 1,000 queries, versus 318.8 ms for IVF-Flat at 99.23%. On DBpedia, both methods miss 0.99 in every fit; HNSW reaches 99.31% mean recall. Flat gives exact strict-ID recovery and the lowest query time on both datasets. IVF-PQ has no feasible selection at any target in the tested grid.

Reference resolution. Figure 4 includes all ten prescribed ψ values at $T = 100$, retaining infeasible band choices. These are validation-screening estimates rather than test-selected operating points.

D.6 Expanded geometric band-count control

Table 10 expands the geometric band grid under the nominal bit capacity of each frozen SMinHash signature. Both families are rebuilt and timed together. The control reuses previously examined test queries and is a post-hoc sensitivity analysis; the capacity bound does not match memory, computation or tuning effort (Appendix C.3.2).

The expanded grid makes geometric LSH validation-feasible at every target. At target 0.95, SMinHash remains more selective on both datasets, but two COYO fits miss the test target while all geometric fits meet it. At target 0.99 on DBpedia, every geometric fit succeeds and every SMinHash fit misses. On COYO, both families meet 0.99 in every fit, with 4.08% candidates for SMinHash versus 54.78% for hyperplanes. Construction reverses the cost advantage: build plus one batch costs 2.984 versus 1.568 seconds. Selectivity alone therefore does not imply lower cost for a newly constructed index.

D.7 Reference resolution and parameter search

Figure 5 varies ψ at fixed $T = 100$ and selects among complete band widths. A larger reference set can reduce both background and true-neighbor collisions, so it need not improve the constrained

Table 10: **More bands restore geometric feasibility while retaining selectivity differences.** Post-hoc sensitivity control: frozen SMinHash settings versus hyperplanes tuned within their nominal bit capacities. Each batch has 1,000 queries and 100k database items on the RTX 5880 Ada GPU. Values are mean$_{\pm \mathrm{SD}}$ over three fits; batch time is the median of three passes including encoding and reranking. † marks any fit missing its validation target on test. Bold marks best complete means within each target, including displayed ties; differing recall prevents an equal-quality interpretation. Nominal bits do not match memory, computation or tuning.

Method	Used bits	Recall@10 (%)	Candidates (%)	Batch (ms)	Complete
DBpedia, target 0.90; Nominal cap 700 bits					
SMinHash	700	$90.16^{\dagger}_{\pm 0.18}$	$\mathbf{1.951}_{\pm 0.020}$	$\mathbf{60.60}_{\pm 0.47}$	3/3
Hyperplane LSH	696	$\mathbf{94.51}_{\pm 0.40}$	$41.527_{\pm 0.223}$	$1092.55_{\pm 13.32}$	3/3
DBpedia, target 0.95; Nominal cap 1000 bits					
SMinHash	1000	$\mathbf{98.81}_{\pm 0.05}$	$\mathbf{4.043}_{\pm 0.011}$	$\mathbf{137.00}_{\pm 1.46}$	3/3
Hyperplane LSH	1000	$98.30_{\pm 0.17}$	$53.136_{\pm 0.588}$	$1399.96_{\pm 2.18}$	3/3
DBpedia, target 0.99; Nominal cap 1000 bits					
SMinHash	1000	$98.81^{\dagger}_{\pm 0.05}$	$\mathbf{4.043}_{\pm 0.011}$	$\mathbf{137.00}_{\pm 1.46}$	3/3
Hyperplane LSH	384	$\mathbf{99.31}_{\pm 0.11}$	$74.405_{\pm 1.095}$	$1947.08_{\pm 21.22}$	3/3
COYO, target 0.90; Nominal cap 800 bits					
SMinHash	800	$91.61_{\pm 0.25}$	$\mathbf{1.266}_{\pm 0.015}$	$\mathbf{47.03}_{\pm 0.68}$	3/3
Hyperplane LSH	512	$\mathbf{93.26}_{\pm 0.40}$	$34.563_{\pm 0.214}$	$939.72_{\pm 3.87}$	3/3
COYO, target 0.95; Nominal cap 700–1000 bits					
SMinHash	700–1000	$\mathbf{96.36}^{\dagger}_{\pm 2.77}$	$\mathbf{3.022}_{\pm 0.925}$	$\mathbf{103.48}_{\pm 42.29}$	3/3
Hyperplane LSH	192–696	$96.09_{\pm 0.77}$	$45.396_{\pm 5.678}$	$1235.49_{\pm 159.10}$	3/3
COYO, target 0.99; Nominal cap 1000 bits					
SMinHash	1000	$\mathbf{99.56}_{\pm 0.01}$	$\mathbf{4.080}_{\pm 0.012}$	$\mathbf{149.36}_{\pm 2.57}$	3/3
Hyperplane LSH	1000	$99.26_{\pm 0.08}$	$54.775_{\pm 0.287}$	$1480.37_{\pm 18.45}$	3/3

operating point. The final selection searches both ψ and b; retrieval quality is evaluated separately on held-out queries.

D.8 Graph and join operating points selected for each task

Table 11 reports all planned graph families under the task-specific protocol of Appendix C.3.4. SMinHash completes every fit on COCO, Deep and SIFT, but exceeds the posting budget on GIST and NYTimes. Geometric LSH is infeasible on NYTimes and exceeds the budget elsewhere. Flat and IVF-Flat complete all five datasets, with lower total cost on the three where SMinHash completes. These outcomes concern the declared grids and limits. FP32 Flat can miss strict neighbor IDs near ties because the reference uses FP64 scores and deterministic tie breaking.

Training-budget and workload sensitivity. Table 12 evaluates the GPU IVF probe and training-budget grid, selecting for projected construction plus one join as described in Appendix C.3.5. All planned full evaluations finish with precision one. Every SMinHash fit meets the recall target and reproduces its original edge IDs and candidate counts. Eleven of twelve one-join-selected IVF fits meet the target; one Deep fit falls below it. IVF's lower total on NYTimes and the nearly equal SIFT totals show that low hash construction cost does not establish a general total-cost advantage. The SIFT difference is small relative to the variation across fits. SMinHash retains the lower mean total on COCO and Deep, with different achieved recalls. Mandatory CPU overflow remains included in IVF costs.

A separate query-only validation objective uses the same grid. Eleven of twelve selected IVF fits again meet the full-join recall target. Its mean query cost is lower than SMinHash on all four datasets, while construction affects the one-join totals. The query-selected configurations need not give the lowest measured full-join query cost: the validation workload and its linear cost projection are selection proxies, not guarantees about the full workload. These are two operating policies for the same IVF family, not two independent algorithms. Identical selected configurations reuse one full measurement. Neither objective includes validation-search cost in the reported post-selection totals.

Table 11: **Complete graph construction after task-specific validation.** Warm is a separately timed full-workload warmup; Query excludes warmup and Total adds construction. HNSW uses eight CPU threads; GPU methods use the RTX 5880 Ada GPU. Inputs are resident on the index device. ANN candidate counts are uninstrumented (–). Done reports complete/planned fits, with NF and C counts retained. Values are mean$_{\pm \mathrm{SD}}$ over three independently fitted seeds; the exact full-range reference has one measured execution and no SD. NF denotes no feasible confirmed validation choice; C is a resource-censored execution. Partial means show completed/expected counts and are excluded from bold winners. Bold marks the best complete displayed mean, including ties, among the shown methods for the same dataset, task and metric. Cost comparisons retain differing achieved recalls. † marks mean recall below the validation target 0.95; individual fits can also miss that target.

Dataset	Method	Recall	Cand. (%)	Warm (s)	Query (s)	Total (s)	Done
COCO512	SMinHash	$0.965_{\pm 0.001}$	$\mathbf{1.10}_{\pm 0.00}$	$2.46_{\pm 1.23}$	$1.77_{\pm 0.07}$	$2.20_{\pm 0.17}$	3/3
	Geo. LSH	C	C	C	C	C	0/3,C3
	IVF-Flat	$0.973_{\pm 0.002}$	–	$\mathbf{0.34}_{\pm 0.00}$	$\mathbf{0.35}_{\pm 0.00}$	$1.35_{\pm 0.06}$	3/3
	HNSW	$0.977_{\pm 0.000}$	–	$2.28_{\pm 0.03}$	$2.32_{\pm 0.12}$	$12.24_{\pm 0.13}$	3/3
	Flat	$\mathbf{1.000}_{\pm 0.000}$	–	$0.59_{\pm 0.00}$	$0.60_{\pm 0.00}$	$\mathbf{0.75}_{\pm 0.01}$	3/3
Deep96	SMinHash	$0.967_{\pm 0.003}$	$\mathbf{3.12}_{\pm 0.02}$	$3.04_{\pm 0.06}$	$2.94_{\pm 0.14}$	$2.98_{\pm 0.14}$	3/3
	Geo. LSH	C	C	C	C	C	0/3,C3
	IVF-Flat	$0.970_{\pm 0.003}$	–	$\mathbf{0.08}_{\pm 0.00}$	$\mathbf{0.08}_{\pm 0.00}$	$0.40_{\pm 0.00}$	3/3
	HNSW	$0.964_{\pm 0.001}$	–	$0.79_{\pm 0.04}$	$0.71_{\pm 0.06}$	$5.40_{\pm 0.21}$	3/3
	Flat	$\mathbf{1.000}_{\pm 0.000}$	–	$0.18_{\pm 0.00}$	$0.18_{\pm 0.00}$	$\mathbf{0.32}_{\pm 0.00}$	3/3
SIFT128	SMinHash	$0.951_{\pm 0.001}$	$\mathbf{1.95}_{\pm 0.02}$	$1.91_{\pm 0.10}$	$1.80_{\pm 0.05}$	$1.88_{\pm 0.04}$	3/3
	Geo. LSH	C	C	C	C	C	0/3,C3
	IVF-Flat	$0.979_{\pm 0.010}$	–	$\mathbf{0.11}_{\pm 0.02}$	$\mathbf{0.11}_{\pm 0.03}$	$0.49_{\pm 0.03}$	3/3
	HNSW	$0.959_{\pm 0.000}$	–	$1.00_{\pm 0.12}$	$1.03_{\pm 0.08}$	$6.53_{\pm 0.13}$	3/3
	Flat	$\mathbf{1.000}_{\pm 0.000}$	–	$0.18_{\pm 0.00}$	$0.18_{\pm 0.00}$	$\mathbf{0.33}_{\pm 0.00}$	3/3
GIST960	SMinHash	C	C	C	C	C	0/3,C3
	Geo. LSH	C	C	C	C	C	0/3,C3
	IVF-Flat	$0.964_{\pm 0.002}$	–	$4.50_{\pm 0.02}$	$4.94_{\pm 0.04}$	$7.09_{\pm 0.02}$	3/3
	HNSW	$0.948^{\dagger}_{\pm 0.000}$	–	$13.89_{\pm 0.71}$	$13.79_{\pm 0.35}$	$35.34_{\pm 1.46}$	3/3
	Flat	$\mathbf{1.000}_{\pm 0.000}$	–	$\mathbf{1.03}_{\pm 0.00}$	$\mathbf{1.05}_{\pm 0.00}$	$\mathbf{1.21}_{\pm 0.01}$	3/3
NYTimes256	SMinHash	C	C	C	C	C	0/3,C3
	Geo. LSH	NF	NF	NF	NF	NF	0/3,NF3
	IVF-Flat	$\mathbf{0.995}_{\pm 0.001}$	–	$3.53_{\pm 0.02}$	$4.06_{\pm 0.03}$	$4.60_{\pm 0.03}$	3/3
	HNSW	$0.949^{\dagger}_{\pm 0.001}$	–	$9.88_{\pm 0.65}$	$9.87_{\pm 0.40}$	$28.82_{\pm 1.52}$	3/3
	Flat	$\mathbf{0.995}_{\pm 0.000}$	–	$\mathbf{0.31}_{\pm 0.00}$	$\mathbf{0.32}_{\pm 0.00}$	$\mathbf{0.47}_{\pm 0.02}$	3/3

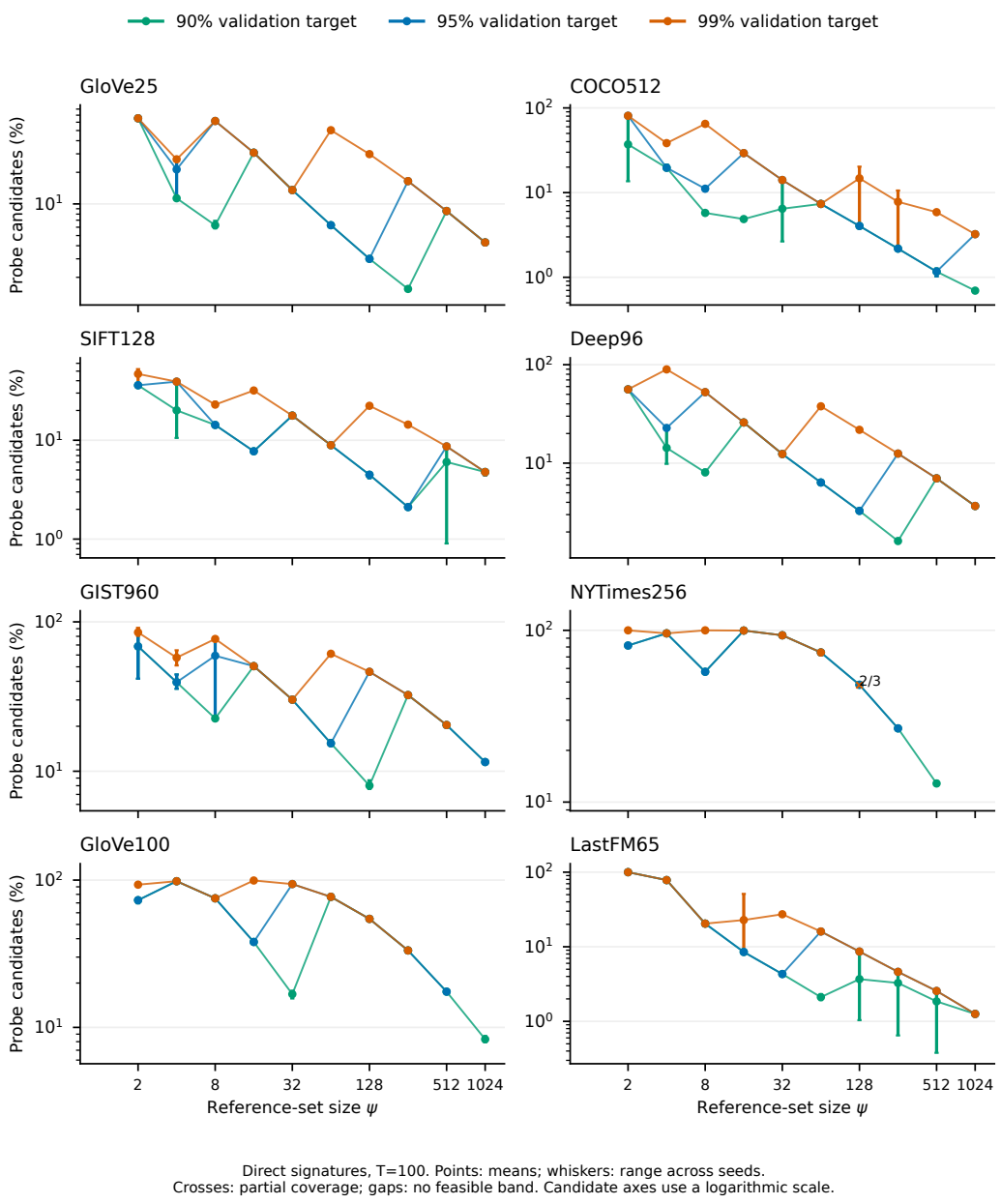

Figure 5: **Reference resolution changes the selectivity–recovery tradeoff.** For each ψ and seed, screening selects the smallest estimated candidate fraction among band widths meeting a validation Recall@10 target. Recall uses exact nearest neighbors of 200 validation queries; candidate fraction uses 8,192 fixed random query–database pairs. Points show means and whiskers span the observed range across three seeds, retaining coverage information where a target is unattainable. The candidate axes are logarithmic. These are validation estimates, not held-out results or query times.

Implementation controls. A CPU-trained and searched IVF control reaches the 0.95 recall target in all twelve fits, with mean recall 97.2–99.7% and precision one. Its query phase is faster than the fixed-batch SMinHash executor on Deep and SIFT, but its training cost is higher. A fixed-training GPU control reaches the target in ten of twelve fits and gives SMinHash lower mean construction-plus-join cost on all four datasets. The expanded grid above contains that GPU control's grid and shows why this total-cost ranking does not persist across training budgets. Execution batches also differ, so the controls do not isolate the effect of moving training to GPU. Their full measurements

Table 12: **Complete-join cost with the construction-plus-join objective.** Mean$_{\pm SD}$ over three fits; query cost is the median of three complete passes after one warmup. Exhaustive has one configuration. All planned full evaluations finish. CPU is the mean percentage of IVF queries requiring the official helper's uncapped overflow search; "–" means this path is not used. GPU IVF construction includes its required CPU clone. Bold marks the highest recall; build, query and total minima require 0.95 recall in every fit. Shading identifies SMinHash. Recall superscripts count fits below 0.95. Both objectives select at validation recall at least 0.95; full achieved recalls differ. Totals exclude tuning. These are shared-host measurements on the reused task cohort.

Dataset	Method	Recall	Build (s)	Query (s)	Total (s)	CPU (%)
COCO512	SMinHash	$0.958^{0}_{\pm 0.004}$	$0.163_{\pm 0.015}$	$0.572_{\pm 0.033}$	$\mathbf{0.734}_{\pm 0.020}$	–
	IVF-GPU†	$0.969^{0}_{\pm 0.003}$	$0.598_{\pm 0.110}$	$\mathbf{0.448}_{\pm 0.075}$	$1.046_{\pm 0.147}$	0.018
	Exhaustive	$\mathbf{1.000}^{0}$	**0.000**	13.697	13.697	–
Deep96	SMinHash	$0.961^{0}_{\pm 0.005}$	$0.064_{\pm 0.007}$	$\mathbf{0.319}_{\pm 0.031}$	$\mathbf{0.383}_{\pm 0.036}$	–
	IVF-GPU†	$0.964^{1}_{\pm 0.029}$	$0.237_{\pm 0.032}$	$0.196_{\pm 0.018}$	$0.434_{\pm 0.021}$	2.183
	Exhaustive	$\mathbf{1.000}^{0}$	**0.000**	3.690	3.690	–
SIFT128	SMinHash	$0.983^{0}_{\pm 0.000}$	$0.038_{\pm 0.004}$	$0.528_{\pm 0.051}$	$\mathbf{0.566}_{\pm 0.049}$	–
	IVF-GPU†	$0.992^{0}_{\pm 0.005}$	$0.252_{\pm 0.006}$	$\mathbf{0.316}_{\pm 0.133}$	$0.567_{\pm 0.128}$	2.088
	Exhaustive	$\mathbf{1.000}^{0}$	**0.000**	5.008	5.008	–
NYTimes256	SMinHash	$0.964^{0}_{\pm 0.001}$	$0.473_{\pm 0.005}$	$0.510_{\pm 0.043}$	$0.984_{\pm 0.048}$	–
	IVF-GPU†	$0.954^{0}_{\pm 0.003}$	$0.388_{\pm 0.039}$	$\mathbf{0.362}_{\pm 0.004}$	$\mathbf{0.750}_{\pm 0.039}$	< 0.001
	Exhaustive	$\mathbf{1.000}^{0}$	**0.000**	7.520	7.520	–

and protocols are retained with the research artifacts. Geometric LSH exceeds its posting budget in all four joins; partial rectangular queries are not substituted for complete joins.

E Distribution and numerical diagnostics

These controls examine the sources of candidate concentration: reference coverage, representation geometry and finite-precision distance evaluation. The final calibration checks distinguish exact collision probabilities from sufficient geometric bounds and their local approximations.

E.1 Isolating the reference source

Primary comparison: independent tuning and new queries. To test adaptation after independent tuning, we give both reference sources the same validation budget: $T = 100$, ten reference sizes $\psi \in \{2, 4, \ldots, 1024\}$, and all nine disjoint band widths dividing 100. For each of three fits, 200 validation queries select the configuration with the fewest complete candidates subject to retaining at least 95% of their exact top-10 neighbors. Ties favor smaller ψ, then larger band width. Each source can select a different configuration; all 48 source/fit selections are fixed before drawing 1,000 new queries per dataset. These query IDs exclude all recorded previous queries and, for the two modern datasets, the canonical 100k database. Some modern query vectors may have appeared as database items in the scale studies; the split establishes new query identities, not semantic independence or never-observed vector content.

Table 13 reports returned Recall@10, complete candidate fractions and complete cost. Candidate retention is computed separately against exhaustive FP64 top-10 truth; it is not used as a substitute for returned recall. Both reference arms score every candidate in FP64. Native GPU Flat and IVF-Flat retain their FP32 inner products and backend tie conventions. IVF parameters come from the original validation-only search over two list counts and five probe counts, minimizing query time subject to 95% recall; the three builds vary clustering initialization without retuning. This provides practical cost context, rather than an identical search space or arithmetic contract.

Construction includes reference extraction or generation, fitting, full database encoding and index construction. Queries and the canonical database are GPU-resident; host result copies are excluded

Table 13: **Reference adaptation improves selectivity, but candidate savings do not determine full query cost.** Independent validation selection of both reference sources precedes a new 1,000-query draw on each 100k database. Values are means with ± sample SD in subscripts over three fits; each query time is the median of eight complete passes. Met counts fits with returned Recall@10 at least 95%. Bold marks the highest recall; bold build and query times identify the lowest cost meeting the target in all three fits; bold candidate fractions compare the two reference sources only when both meet it in all fits. Shading marks SMinHash with data references. Native ANN uses FP32 scoring, whereas both reference arms use FP64 candidate scoring.

Dataset	Family	Met	Recall@10 (%)	Candidates (%)	Build (s)	Query (ms)
Deep96	SMinHash	3/3	$96.607_{\pm 0.093}$	$\mathbf{3.146}_{\pm 0.009}$	$0.864_{\pm 1.290}$	$72.218_{\pm 16.998}$
	Sphere	3/3	$97.500_{\pm 0.164}$	$12.094_{\pm 0.095}$	$\mathbf{0.166}_{\pm 0.023}$	$62.779_{\pm 20.466}$
	Flat	3/3	$\mathbf{100.000}_{\pm 0.000}$	–	$0.397_{\pm 0.153}$	$13.673_{\pm 4.928}$
	IVF-Flat	3/3	$97.110_{\pm 0.078}$	–	$2.172_{\pm 0.280}$	$\mathbf{7.758}_{\pm 1.075}$
COCO512	SMinHash	2/3	$95.767_{\pm 1.547}$	$0.922_{\pm 0.291}$	$1.035_{\pm 1.186}$	$83.595_{\pm 7.764}$
	Sphere	3/3	$97.523_{\pm 0.926}$	$30.908_{\pm 4.432}$	$0.939_{\pm 0.204}$	$209.738_{\pm 14.649}$
	Flat	3/3	$\mathbf{100.000}_{\pm 0.000}$	–	$\mathbf{0.275}_{\pm 0.118}$	$21.496_{\pm 4.149}$
	IVF-Flat	3/3	$97.440_{\pm 0.316}$	–	$3.587_{\pm 0.461}$	$\mathbf{5.887}_{\pm 0.389}$
GloVe25	SMinHash	3/3	$95.890_{\pm 0.210}$	$\mathbf{3.001}_{\pm 0.021}$	$\mathbf{0.118}_{\pm 0.021}$	$91.399_{\pm 22.790}$
	Sphere	3/3	$96.970_{\pm 0.147}$	$6.006_{\pm 0.111}$	$0.140_{\pm 0.030}$	$94.098_{\pm 16.501}$
	Flat	3/3	$\mathbf{100.000}_{\pm 0.000}$	–	$0.342_{\pm 0.210}$	$\mathbf{13.768}_{\pm 13.792}$
	IVF-Flat	3/3	$95.347_{\pm 0.091}$	–	$1.695_{\pm 0.338}$	$15.616_{\pm 7.609}$
GloVe100	SMinHash	3/3	$98.223_{\pm 0.156}$	$\mathbf{17.056}_{\pm 0.018}$	$\mathbf{0.149}_{\pm 0.075}$	$66.008_{\pm 17.741}$
	Sphere	3/3	$97.610_{\pm 0.190}$	$31.323_{\pm 1.181}$	$0.227_{\pm 0.061}$	$78.331_{\pm 19.857}$
	Flat	3/3	$\mathbf{100.000}_{\pm 0.000}$	–	$0.230_{\pm 0.048}$	$\mathbf{7.368}_{\pm 1.457}$
	IVF-Flat	3/3	$98.140_{\pm 0.213}$	–	$1.755_{\pm 0.620}$	$10.255_{\pm 8.051}$
LastFM65	SMinHash	3/3	$99.743_{\pm 0.093}$	$\mathbf{1.417}_{\pm 0.021}$	$0.193_{\pm 0.026}$	$73.275_{\pm 29.208}$
	Sphere	3/3	$98.883_{\pm 0.272}$	$94.221_{\pm 0.558}$	$\mathbf{0.133}_{\pm 0.039}$	$293.548_{\pm 15.471}$
	Flat	3/3	$\mathbf{100.000}_{\pm 0.000}$	–	$0.204_{\pm 0.105}$	$\mathbf{7.032}_{\pm 2.050}$
	IVF-Flat	3/3	$98.450_{\pm 0.168}$	–	$1.275_{\pm 0.422}$	$8.729_{\pm 4.950}$
NYTimes256	SMinHash	3/3	$97.327_{\pm 0.096}$	$\mathbf{26.467}_{\pm 0.045}$	$\mathbf{0.185}_{\pm 0.031}$	$107.656_{\pm 6.808}$
	Sphere	3/3	$98.570_{\pm 0.235}$	$52.458_{\pm 0.078}$	$0.404_{\pm 0.182}$	$179.948_{\pm 7.018}$
	Flat	3/3	$99.420_{\pm 0.000}$	–	$0.444_{\pm 0.107}$	$\mathbf{11.462}_{\pm 2.103}$
	IVF-Flat	3/3	$\mathbf{99.430}_{\pm 0.030}$	–	$2.528_{\pm 0.442}$	$30.323_{\pm 2.065}$
DBpedia1536	SMinHash	3/3	$98.893_{\pm 0.064}$	$\mathbf{3.995}_{\pm 0.007}$	$0.619_{\pm 0.158}$	$114.371_{\pm 5.290}$
	Sphere	3/3	$96.360_{\pm 2.252}$	$32.694_{\pm 11.252}$	$1.632_{\pm 1.005}$	$671.209_{\pm 224.684}$
	Flat	3/3	$\mathbf{100.000}_{\pm 0.000}$	–	$\mathbf{0.423}_{\pm 0.045}$	$\mathbf{33.870}_{\pm 2.960}$
	IVF-Flat	0/3	$94.450_{\pm 0.079}$	–	$6.695_{\pm 0.369}$	$48.120_{\pm 2.744}$
COYO2048	SMinHash	1/3	$95.027_{\pm 0.193}$	$2.522_{\pm 0.024}$	$0.515_{\pm 0.150}$	$87.156_{\pm 8.483}$
	Sphere	3/3	$97.417_{\pm 0.091}$	$28.582_{\pm 0.450}$	$2.905_{\pm 0.164}$	$606.529_{\pm 10.095}$
	Flat	3/3	$\mathbf{100.000}_{\pm 0.000}$	–	$\mathbf{0.282}_{\pm 0.039}$	$\mathbf{45.419}_{\pm 1.617}$
	IVF-Flat	0/3	$94.920_{\pm 0.070}$	–	$7.289_{\pm 1.265}$	$39.141_{\pm 3.762}$

for all families. Each complete query pass includes encoding, bucket lookup, candidate union, scoring and output assembly in requests of 32, with an eight-query tail. We warm the actual request shapes and rotate the family order over eight timed passes. The table reports means and sample standard deviations across three fits of their median pass times. The GPU model is unchanged; each dataset/fit uses one of two monitored devices, without claiming exclusive-device access. All planned primary fits complete within the declared resource limits.

Data references yield lower mean candidate fractions on 8 of eight datasets. Both sources reach the returned 95% target in 21 of 24 paired fits, and data references pass fewer candidates in all 21 of these pairs. The misses remain in the table: data references meet the target in two of three COCO fits and one of three COYO fits. Thus equal validation targets do not imply equal achieved quality on

new queries. The spherical control meets the target in all fits; IVF-Flat misses it on DBpedia and COYO in all three fits.

Lower candidate volume translates into lower mean complete query time than the spherical control on 7 of eight datasets. Deep is the exception: the data arm uses a larger reference set, and fewer candidates do not offset its complete processing cost. Data-reference construction is cheaper than IVF-Flat in all 8 dataset means, but native Flat queries are faster in all 8. The experiment therefore supports a conditional adaptation advantage while retaining the cost limits of the minimum-index candidate pipeline. The 90% and 99% sensitivity settings retain their validation-selected parameters; their complete candidate diagnostics are included in the reproducibility artifacts rather than additional per-fit manuscript tables.

Fixed-parameter reference-source intervention. Data-derived references can adapt to populated regions, an intuition also underlying Kang & Jung (2012). We test the effect of the reference source within the same nearest-index rule. For each of the eight 100k-item datasets and three fits, we pair the original 16,384-item data reference pool with 16,384 independent uniform unit vectors in the same ambient dimension. The control vectors are normalized Gaussian draws, generated independently of the data. Both arms use exactly the same ordered reference indices, $T = 100$, previously selected ψ, first-minimum tie rule and contiguous bands. Across the three fits, $\psi = 1024$ on COCO, GloVe25, LastFM and COYO, and $\psi = 512$ on DBpedia; the ranges are 512–1024 on Deep, 128–256 on GloVe100, and 16–64 on NYTimes. Both are newly encoded through the same FP32 distance and reduction path, with TF32 disabled. Thus the intervention changes reference coordinates while preserving the sampling map and encoder. The parameters were selected for data references; this is a fixed-parameter ablation, not an independently tuned comparison. The 1,000 queries per dataset are the existing query set, so this analysis is not a new test-set confirmation. Re-encoding changes fewer than one in a million coordinates relative to the archived data-reference codes; all per-query top-10 retention counts for the three band widths remain unchanged.

Table 14 asks whether replacing the reference source changes bucket concentration while retaining useful neighbors. At $b = 1$, data references lower both load measures on seven of eight datasets. On COCO512, the largest-bucket fraction falls from 37.754% to 1.226%, while top-10 candidate retention changes from 99.967% to 99.883%. NYTimes is the exception: data references have larger maximum buckets and a slightly higher mean pair fraction. Across datasets, data-reference retention remains at least 99.460%; the spherical control has higher retention on COCO512, GloVe25 and GloVe100. The LastFM disparity is consistent with the asymmetric representation examined in Appendix E.3.

The same fixed intervention also evaluates $b = 2$ and $b = 5$. Data references retain more top-10 neighbors on all eight datasets at both band widths. At $b = 2$, data references lower the maximum-bucket statistic on seven datasets but the mean pair fraction on three; at $b = 5$, these counts fall to three and one. Their load advantage is therefore less uniform after concatenation. Band work depends on higher moments of pair-specific collision probabilities (Equation (77)). These observations identify a useful reference-source effect against this isotropic control. They establish neither a distribution-free balance guarantee nor a runtime advantage at matched recovery.

E.2 Measured bucket loads at matched key capacity

We examine the frozen 100k-item database codes from the eight-dataset study, using its three fits and validation-selected ψ values. For each fit, group the first $100 \log_2 \psi$ Hyperplane1000 bits into 100 categorical coordinates. Each control coordinate then has the same nominal alphabet size ψ as a SMinHash coordinate. Contiguous bands of width $b \in \{1, 2, 5\}$ give both methods $100/b$ tables with nominal key capacity ψ^b. No bits or configurations are selected using the occupancy results. This is a post-hoc comparison at matched nominal capacity, not matched recall or a causal isolation of data dependence.

For each fit we measure the largest bucket across tables, divided by N, and the mean per-table directed non-self collision fraction

$$\overline{c}_b = \frac{1}{LN(N-1)} \sum_{\ell,g} n_{\ell g}(n_{\ell g} - 1).$$

Table 14: **Changing the reference source exposes distribution-sensitive loads.** Fixed nearest-index rule, $T = 100$, $b = 1$, and validation-selected ψ; 100k items and 1,000 existing queries. Data denotes SMinHash with data references; Sphere replaces only the reference bank with independent uniform unit vectors. Values are percentages, mean$_{\pm \mathrm{SD}}$ over three paired fits. Shading identifies data references; bold marks the better displayed mean within each pair (lower load, higher retention). Retention concerns top-10 candidates, not reranked recall. This intervention does not retune the control.

Dataset	References	Largest bucket ↓	Pair fraction ↓	Retention ↑
Deep96	Data	$\mathbf{1.639}_{\pm 0.517}$	$\mathbf{0.259}_{\pm 0.086}$	$\mathbf{99.783}_{\pm 0.231}$
	Sphere	$4.883_{\pm 0.120}$	$0.491_{\pm 0.110}$	$99.483_{\pm 0.670}$
COCO512	Data	$\mathbf{1.226}_{\pm 0.098}$	$\mathbf{0.199}_{\pm 0.001}$	$99.883_{\pm 0.035}$
	Sphere	$37.754_{\pm 16.007}$	$4.470_{\pm 0.660}$	$\mathbf{99.967}_{\pm 0.015}$
GloVe25	Data	$\mathbf{0.933}_{\pm 0.105}$	$\mathbf{0.149}_{\pm 0.000}$	$99.550_{\pm 0.052}$
	Sphere	$6.390_{\pm 0.532}$	$0.555_{\pm 0.023}$	$\mathbf{99.907}_{\pm 0.057}$
GloVe100	Data	$\mathbf{10.320}_{\pm 3.441}$	$\mathbf{1.287}_{\pm 0.425}$	$99.827_{\pm 0.179}$
	Sphere	$14.197_{\pm 4.477}$	$1.550_{\pm 0.444}$	$\mathbf{99.887}_{\pm 0.153}$
LastFM65	Data	$\mathbf{2.120}_{\pm 0.171}$	$\mathbf{0.179}_{\pm 0.001}$	$\mathbf{99.720}_{\pm 0.095}$
	Sphere	$99.999_{\pm 0.001}$	$87.911_{\pm 3.867}$	$7.357_{\pm 1.780}$
NYTimes256	Data	$10.913_{\pm 5.192}$	$5.008_{\pm 2.828}$	$\mathbf{99.970}_{\pm 0.052}$
	Sphere	$\mathbf{9.122}_{\pm 4.363}$	$\mathbf{4.954}_{\pm 2.809}$	$99.940_{\pm 0.104}$
DBpedia1536	Data	$\mathbf{1.081}_{\pm 0.024}$	$\mathbf{0.240}_{\pm 0.001}$	$\mathbf{99.617}_{\pm 0.025}$
	Sphere	$3.584_{\pm 0.656}$	$0.455_{\pm 0.004}$	$94.983_{\pm 0.357}$
COYO2048	Data	$\mathbf{0.816}_{\pm 0.038}$	$\mathbf{0.144}_{\pm 0.000}$	$\mathbf{99.460}_{\pm 0.066}$
	Sphere	$2.983_{\pm 1.095}$	$0.303_{\pm 0.010}$	$89.037_{\pm 0.280}$

The latter is the probability that a uniform distinct database pair shares a key in a uniformly chosen table. It includes repeated matches across tables and is not a distinct candidate union or a query-time measurement. We compare means over the three fits, with $b = 1$ as the primary control. SMinHash lowers both statistics on six of eight datasets, with the largest difference on LastFM; GloVe100 and NYTimes have slightly larger mean pair fractions. The LastFM representation effect is examined separately in Appendix E.3.

The advantage need not persist under concatenation. With $b = 2$, SMinHash has the smaller largest-bucket statistic on three datasets and smaller mean pair fraction on two; with $b = 5$, it lowers both only on LastFM. Equation (77) explains why a smaller average single-round collision does not imply smaller band loads: banding depends on the higher moments of pair-specific collision probabilities, $\mathbb{E}_{X,Y}[p_\psi(X, Y)^b]$.

To interpret this change, a secondary post-hoc check applies the same fixed bands to all 1,000 existing queries and their numerical top-10 neighbors. SMinHash retains more neighbors on all eight datasets at both $b = 2$ and $b = 5$. At $b = 2$, its dataset means range from 70.603% to 93.497%, versus 0.013% to 62.060% for the grouped hyperplane control. Neither range establishes the primary 95% target. At $b = 5$, both methods lose substantial coverage: the corresponding ranges are 6.880–32.437% and 0–3.510%. These are candidate-retention checks, not reranked retrieval results or new independent query draws. Larger buckets can retain useful neighbor structure, so occupancy and recovery must be considered together. The observations support distribution-dependent load improvements; they do not imply that reference sampling prevents hot buckets or that lower bucket work at unmatched recovery gives a faster useful index.

E.3 Reference selection under asymmetric user–item geometry

The large LastFM screening difference has a representation-specific interpretation. The publisher converts inner-product search to angular search by equalizing item norms with an extra coordinate and appending zero to user queries.[3] We retain these published vectors. The analysis below describes this input geometry; it introduces no change to the encoder or candidate rule. The lift is the construction

[3]The publisher's `lastfm` generator: `https://github.com/erikbern/ann-benchmarks/blob/main/ann_benchmarks/datasets.py`.

used by SIMPLE-LSH (Neyshabur & Srebro, 2015). Normalization-induced concentration and its effect on sign hashing were already analyzed by Yan et al. (2018), who proposed partitioning by item norms. We do not claim that phenomenon as new; the comparison below concerns reference-index assignments on the fixed representation.

An exact-arithmetic comparison. Let a finite collection of original items satisfy $\|u\| \le M$, choose $R > M$, and define unit vectors

$$P_R(u) = \left(\frac{u}{R}, \sqrt{1 - \frac{\|u\|^2}{R^2}}\right), \qquad Q(q) = \left(\frac{q}{\|q\|}, 0\right), \quad q \neq 0. \tag{94}$$

Their inner product is $\langle q, u\rangle/(R\|q\|)$, so changing R preserves the original query–item ranking. Fix the same ordered reference samples $(v_0, \dots, v_{\psi-1})$ for every R. A query's selected index is

$$\arg\min_j \|Q(q) - P_R(v_j)\|^2 = \operatorname*{argmax}_j \langle q, v_j\rangle, \tag{95}$$

including the shared local-index tie rule. Thus the query signature is independent of R in real arithmetic.

For a database item and reference, rationalizing the square roots gives

$$\begin{aligned} R^2\|P_R(u) - P_R(v)\|^2 &= \|u - v\|^2 + \frac{(\|v\|^2 - \|u\|^2)^2}{(\sqrt{R^2 - \|u\|^2} + \sqrt{R^2 - \|v\|^2})^2}, \\ 0 \le R^2\|P_R(u) - P_R(v)\|^2 &- \|u - v\|^2 \le \frac{M^4}{4(R^2 - M^2)}. \end{aligned} \tag{96}$$

Consequently, a unique Euclidean reference winner with squared-distance gap $\Delta > 0$ remains the winner once the right-hand bound is below Δ. The database assignments approach nearest-reference assignments in the original item space, provided these gaps are positive.

In contrast, a fixed hyperplane $w = (a, b)$ has

$$\langle w, P_R(u)\rangle = \frac{\langle a, u\rangle}{R} + b\sqrt{1 - \frac{\|u\|^2}{R^2}}. \tag{97}$$

If $b \neq 0$ and $R > M\sqrt{1 + \|a\|^2/b^2}$, the second term's magnitude exceeds the first for every item. All items then have sign $\operatorname{sign}(b)$. For finitely many Gaussian hyperplanes, every b is nonzero almost surely, so a finite common R makes all database signatures identical. A fixed query's complete count gate then accepts all items or none. With exactly k true neighbors per query, candidate fraction equals neighbor retention in this limit. This conclusion concerns hyperplanes on the stated representation, not methods with another inner-product-search transformation.

A finite construction with retained matches. The preceding limit alone does not guarantee reference agreement for correct pairs. The following construction gives an explicit sufficient condition and an exact false-collision rate.

Proposition E.1 (Finite selectivity separation)**.** *Let $K \ge 2$, $R > 1$, $x_i = P_R(e_i)$ and $q_i = Q(e_i)$, where $e_1, \dots, e_K$ are the standard basis of $\mathbb{R}^K$. Independently each round, sample an ordered tuple of $1 \le \psi \le K$ distinct database items uniformly, using the shared first-index tie rule. In real arithmetic, x_i is the unique nearest item to q_i and $h(q_i) = h(x_i)$ for every sample. For $i \neq j$,*

$$\Pr\{h(q_i) = h(x_j)\} = p_\psi = \frac{(K - \psi)(K - \psi + 1)}{K(K - 1)}. \tag{98}$$

With $T = Lb$ independent rounds, the complete band rule retains every correct pair and has expected candidate fraction

$$\frac{1 + (K - 1)\{1 - (1 - p_\psi^b)^L\}}{K}. \tag{99}$$

For a complete count gate, replace the braces by $\Pr\{\mathrm{Bin}(T, p_\psi) \ge m\}$, where $1 \le m \le T$. In particular, $\psi = K$ gives retention one and candidate fraction $1/K$. For H Gaussian hyperplanes on the same fixed inputs, all database signatures are identical with probability at least

$$\max\left\{0,\, 1 - \frac{HK}{\pi}\arcsin\frac{1}{R}\right\}. \tag{100}$$

On this event, any complete band or count gate admits all items or none per query; averaged over the K queries, its candidate fraction equals correct-pair retention.

Proof. The query–item inner product is $1/R$ for the matched item and zero otherwise. Query q_i and item x_i both choose the local position of i if it is sampled, and the first position otherwise. For distinct i, j, a collision occurs exactly when neither is sampled, or when one is first and the other is absent. These disjoint events have total probability

$$\frac{(K-\psi)(K-\psi-1)+2(K-\psi)}{K(K-1)},$$

which gives Equation (98), including $p_1 = 1$ and $p_K = 0$. Independence across rounds gives the band and count expressions; linearity of expectation over items gives Equation (99), without independence between different item pairs.

For a hyperplane (a, β), the angle between x_i and the last-coordinate axis is $\arcsin(1/R)$. Rotational symmetry of the two Gaussian coordinates (a_i, β) makes the probability of a sign different from $\operatorname{sign}(\beta)$ equal to $\arcsin(1/R)/\pi$. A union bound over HK item–plane pairs gives Equation (100). On its event, all items have one common signature. Admitting a query's correct item therefore admits the whole database, proving the final statement. □

This is a selectivity separation for paired identity queries and an equal-norm reference pool, not a timing comparison or an assumption about LastFM's original factors. It keeps the encoder unchanged. At large R, finite precision can destroy the required distance ordering; even summation roundoff can disturb the constructed ties. The proposition consequently makes no bitwise floating-point guarantee.

Measurements on the unchanged LastFM inputs. In our 100k-item database, the appended coordinate has mean 0.997 and sample standard deviation 0.013; it is exactly zero for all 1,000 test queries. Hyperplane1000 has 445–480 constant database coordinates across three fits, compared with none of SMinHash's 100 coordinates. The three-fit mean pooled AUC for ranking numerical top-10 pairs above non-neighbor pairs by agreement count is 99.783% for SMinHash and 90.330% for Hyperplane1000, counting ties as one half. Negative-pair histograms exclude the 10 true neighbors of every query. The corresponding candidate fractions and achieved retention are reported in Table 4. On NYTimes the ranking reverses: the same diagnostic gives 97.941% versus 99.845%. These are post-hoc descriptions of the existing fixed test split, without threshold reselection or a new independent query draw.

A fixed origin control. To isolate the role of the origin, we subtract the mean of the fixed 16,384-item reference pool from both database and query vectors before applying each of the same three 1,000-plane matrices. The center uses no query labels. For each representation, 200 validation queries select the largest agreement threshold retaining at least 95% of top-10 pairs; the same 1,000 test queries then measure complete candidate admission. Both sign controls use FP64 arithmetic on the saved FP32 inputs and planes. The raw control differs from the earlier FP32 codes in two of 300 million database bits and no test-query bits. This is a post-hoc control on an already observed test split, not an independent replication or a change to SMinHash.

For the squared-Euclidean reference rule, a common translation preserves every index in real arithmetic, including ties, since

$$\|(x-c)-(z_j-c)\|^2 = \|x-z_j\|^2 \qquad \text{for every } j.$$

The sign rule instead compares $w^\top x$ with $w^\top c$ after translation. This identity explains the different origin dependence without rerunning or modifying SMinHash. It does not assert invariance after separately renormalizing the translated vectors or under floating-point rounding.

Centering removes every constant database coordinate and reduces the sign control's mean candidate volume by 61.924×. Its achieved retention is 94.830%, below the validation target in all three fits. SMinHash retains 95.823% with fewer observed candidates, but these operating points are not a recall-matched comparison. The control confirms that the original 136-fold gap is strongly representation-dependent. Centering also changes angular geometry; this experiment is neither a new cosine-LSH guarantee nor an evaluation of specialized inner-product-search methods. No timing conclusion follows from these CPU candidate measurements.

Table 15: **The origin explains much of the LastFM sign-hash gap.** Complete candidate admission over 100k items and 1,000 fixed test queries; mean$_{\pm \mathrm{SD}}$ over three fits. Raw and centered controls use the same planes and their own validation-selected thresholds. The unchanged SMinHash result is included for context and shaded. Bold marks the highest retention mean; candidate minima require 95% mean retention. These are candidate retention measurements, not returned-neighbor recall or query timings.

Encoder	Retention (%)	Candidates (%)	Fits $\geq 95\%$
SMinHash (unchanged)	$\mathbf{95.823}_{\pm 0.516}$	$\mathbf{0.332}_{\pm 0.027}$	3/3
Hyperplane1000, raw	$95.263_{\pm 0.739}$	$45.298_{\pm 3.606}$	2/3
Hyperplane1000, centered	$94.830_{\pm 0.026}$	$0.732_{\pm 0.087}$	0/3

The general limit explains a possible discrimination loss, while the finite-input measurements establish the observed behavior. Outside the construction's stated conditions, these arguments do not guarantee that queries and their true neighbors select the same reference. Nor is Equation (95) a bitwise FP32 guarantee: sufficiently large scaling can also erase nearest-reference score differences numerically. Overall AUC does not establish recovery at a validation-selected threshold, and neither AUC nor candidate volume alone establishes a complete-query speed advantage.

E.4 Reference adaptation under a common representation component

The reference-source intervention above changes an observed index's load, but does not isolate why. Here we examine one mechanism: references that co-move with the data preserve its relative geometry, whereas references fixed in the ambient coordinates can acquire a common preferred winner. The elementary invariance below is shared by other distance-based procedures. It is an explanation of the encoder's behavior, not a claim that data dependence guarantees balance or that the invariance is new.

Transported references preserve the complete signature. For $d \geq 1$, let $Q : \mathbb{R}^d \to \mathbb{R}^{d'}$ be a linear isometry, $a \in \mathbb{R}^{d'}$, and $s > 0$. Apply $A(v) = sQv + a$ to every input and reference, and couple the ordered sampled reference indices. For every x, z,

$$\|A(x) - A(z)\|^2 = s^2\|x - z\|^2, \qquad h_{A(Z_t)}(A(x)) = h_{Z_t}(x). \tag{101}$$

The first identity follows from $Q^\top Q = I$; multiplication by the same positive scalar preserves all comparisons and ties, proving the second identity for the first-local-index rule. Every coordinate, band key, count gate and bucket size is therefore unchanged in exact arithmetic. This holds for any fixed reference source, provided it is transported as well. Existing hot buckets remain hot.

To introduce a common component, embed x as $\bar{x} = (x, 0)$ and set

$$A_\varepsilon(v) = \varepsilon v + \sqrt{1 - \varepsilon^2}\, e_{d+1}, \qquad 0 < \varepsilon \leq 1. \tag{102}$$

For unit x, $A_\varepsilon(\bar{x})$ remains unit length, with $\langle A_\varepsilon(\bar{x}), A_\varepsilon(\bar{y})\rangle = 1 - \varepsilon^2 + \varepsilon^2\langle x, y\rangle$. Nearest-neighbor ranks are preserved. A Euclidean threshold must scale by ε, and a cosine threshold τ becomes $1 - \varepsilon^2 + \varepsilon^2\tau$; absolute distances are not fixed. The zero endpoint, where every input coincides, is excluded.

A fixed external reference set can collapse uniformly. For one round, write fixed references as $w_j = (u_j, v_j)$, and suppose v_* is their unique largest last coordinate. For all original inputs satisfying $\|x\| \leq R$, define

$$\gamma_\varepsilon = \min_{j \neq *} \left\{ \|w_j\|^2 - \|w_*\|^2 + 2\sqrt{1 - \varepsilon^2}(v_* - v_j) - 2\varepsilon R\|u_* - u_j\| \right\}. \tag{103}$$

If $\gamma_\varepsilon > 0$, every such input selects w_*. Indeed, squared-distance differences satisfy

$$\begin{aligned} &\|A_\varepsilon(\bar{x}) - w_j\|^2 - \|A_\varepsilon(\bar{x}) - w_*\|^2 \\ =&\|w_j\|^2 - \|w_*\|^2 + 2\sqrt{1 - \varepsilon^2}(v_* - v_j) + 2\varepsilon\langle x, u_* - u_j\rangle \;\geq\; \gamma_\varepsilon, \end{aligned}$$

where the inequality uses Cauchy–Schwarz and $\|x\| \le R$. For unit references the norm difference vanishes. With $\Delta = \min_{j \neq *}(v_* - v_j) > 0$ and $R = 1$, the simpler sufficient condition

$$\sqrt{1-\varepsilon^2}\,\Delta > 2\varepsilon, \quad \text{equivalently} \quad \varepsilon < \frac{\Delta}{\sqrt{4+\Delta^2}},$$

uses $\|u_* - u_j\| \le 2$. A finite bank of independent uniform sphere references has distinct last coordinates almost surely. For any finite number of sampled rounds, all rounds therefore collapse for sufficiently small positive ε; every band then contains the entire database in one bucket. Independence across rounds is unnecessary for this conclusion. This statement concerns the specified fixed sphere control. Transporting that control removes this mechanism, and other data-independent families need not behave the same way.

Paired intervention and numerical scope. We use all eight 100k-item datasets, three fits, the same ordered reference indices, $T = 100$, and the previously selected ψ values. For $\varepsilon \in \{1, 0.3, 0.1, 0.03, 0.01, 0.003\}$, we compare transported data references, fixed uniform sphere references in $d+1$ dimensions, and the same sphere references transported by A_ε. The external bank contains 16,384 references per fit. The last arm generally has non-unit reference norms; the unchanged squared-distance encoder includes these norms, with no renormalization. All three arms use the same FP32 implementation with TF32 disabled. The fixed and transported sphere arms are identical at $\varepsilon = 1$. Parameters are not retuned in this diagnostic.

The original 1,000 queries and top-10 target IDs are held fixed, so their retention measures preservation of the original targets, not a new test-set result. The similarity identity applies to the stored vectors in real arithmetic; it does not promise unchanged FP32 codes. On a fixed subset of eight database and eight query rows per fit, we also evaluate all rounds with direct FP64 squared distances, both before and after rounding transformed inputs to FP32. This distinguishes input rounding from the production norm-expansion and reduction path.

Figure 6 isolates the effect of reference co-movement. At $\varepsilon = 0.01$, every fixed-sphere fit has a band containing all 100,000 database items; consequently every database pair is a candidate. At $\varepsilon = 0.03$, data-reference coordinate changes are only 0.047–0.098% on seven datasets, with mean target-retention changes below 0.014 percentage points. LastFM is the exception discussed below. Transported sphere references also retain their original load patterns, including the already concentrated LastFM baseline. This control attributes the geometric stability to reference co-movement; it does not imply that either reference source balances every dataset.

The same intervention exposes a numerical limit. At $\varepsilon = 0.003$, LastFM's data-reference within-band pair fraction rises from 0.006% to 86.727%, while original-target retention falls from 17.687% to 1.977%. Thus the current FP32 implementation can itself produce severe concentration. Across the transported-reference arms, none of the 460,800 sampled FP64 coordinate evaluations changes from its baseline; rounding inputs before direct FP64 distance evaluation changes only two coordinates. The much larger production discrepancies therefore arise primarily within the FP32 encoding path. These observations limit an implementation-level invariance claim, even though the exact-arithmetic identity remains valid. The intervention establishes neither universal balance nor an advantage over appropriately adapted alternative indexes.

E.5 Separating geometric concentration from numerical collapse

The preceding intervention exposes a difference between an invariant nearest-reference rule and its FP32 evaluation. When inputs and references share a large component, norm expansion can lose the small differences that determine the winning index. We evaluate a standard numerical remedy: subtract the same origin from inputs and references before computing distances. Centering to reduce this cancellation is also discussed in the Faiss documentation (Faiss contributors, 2025); it is not a new hashing rule or a novelty claim.

Equivalent scores and a conditional error certificate. For a fixed origin c, let $u = x - c$ and $v_j = z_j - c$. The score

$$s_j(c) = \|v_j\|^2 - 2\langle u, v_j\rangle = \|x - z_j\|^2 - \|u\|^2 \tag{104}$$

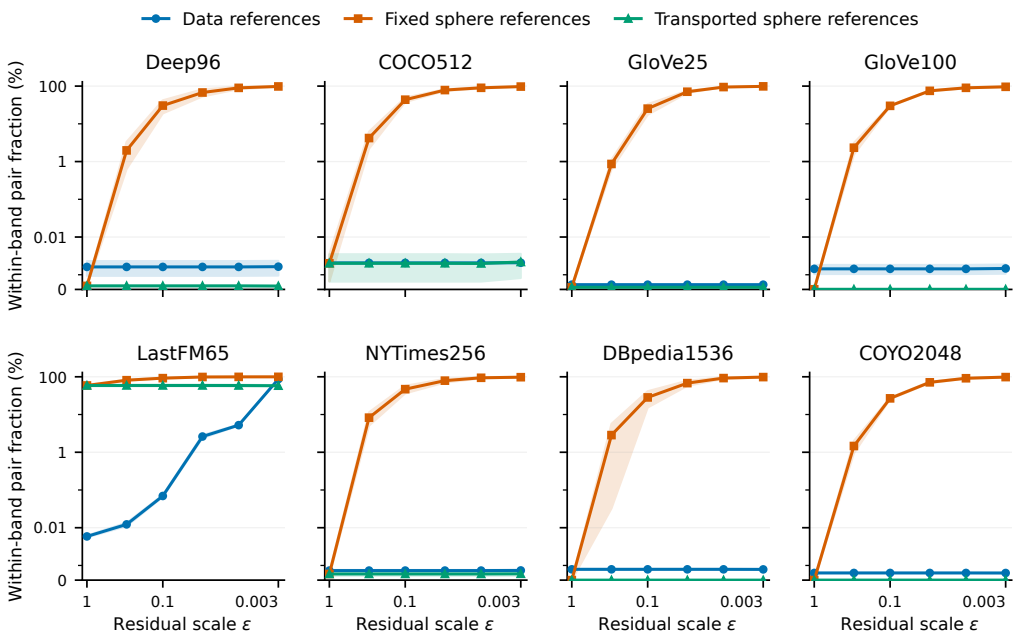

Figure 6: **Reference co-movement and within-band candidate work.** Directed non-self database pair fraction sharing a band, averaged over the 20 bands of a $T = 100$, $b = 5$ signature. Curves and shaded regions show the mean and one sample SD over three fits for each dataset. Decreasing ε strengthens the common component while preserving nearest-neighbor ranks in exact arithmetic. The vertical axis is logarithmic above 0.001% and linear near zero. All reference arms use the same FP32 nearest-index encoder and ordered samples. This fixed-parameter intervention measures candidate work, not runtime or retrieval quality at matched recovery.

has exactly the same first-local argmin as the original squared distance. The subtracted term is independent of j, so all ties are preserved in real arithmetic. There is no renormalization after translation.

For represented inputs, errors from translation and score evaluation can still change a nearly tied minimum. If each computed score has absolute error at most E_j relative to a common-offset exact score, a unique winner w is preserved whenever

$$d_j^2 - d_w^2 > E_j + E_w \qquad \text{for every } j \neq w.$$

This follows by subtracting the two error intervals; it is a conditional certificate, not a guarantee that FP32 resolves every margin. Standard rounding bounds scale with $\|v_j\|^2 + 2\langle |u|, |v_j| \rangle$ (Higham & Mary, 2019). Removing a large shared component can reduce this scale, while input conversion contributes its own error. The diagnostic below checks represented-input argmins directly.

Paired numerical evaluation. We repeat all eight datasets, three fits, six nonzero scales and three reference sources from Appendix E.4, and add the original unpadded data-reference representation. These give 456 geometry conditions evaluated through two numerical paths. Each uses the same ordered reference indices, $T = 100$, selected ψ, first-local minimum and candidate gates. The translated path uses the FP64 mean of the full 16,384-item reference bank as c, subtracts it in FP64, and converts the result to FP32 for the same matrix multiplication and index reduction. TF32 is disabled in both paths. The origin is fixed by the bank; queries and their targets do not select it.

Full database and query signatures determine bucket loads and retention of the original top-10 targets. Direct FP64 squared distances on the same eight database and eight query rows per fit diagnose index errors relative to the represented FP32 inputs. The existing queries and parameters are reused; this is a numerical ablation, not a new test-set confirmation. All task comparisons elsewhere retain their original arithmetic and measured results.

Table 16: **Equivalent arithmetic can prevent numerically induced bucket concentration.** Data-derived references at $\varepsilon = 0.003$, $T = 100$, $b = 5$; 100k database items and 1,000 existing queries. Raw and Translated use the same reference maps and candidate rule. Values are percentages, mean$_{\pm SD}$ over three fits. Retention concerns the original top-10 targets, not returned recall. Shading marks translated arithmetic; bold marks the better displayed mean within each pair.

	Largest bucket ↓		Target retention ↑	
Dataset	Raw	Translated	Raw	Translated
Deep96	$\mathbf{0.200}_{\pm 0.054}$	$\mathbf{0.200}_{\pm 0.060}$	$\mathbf{21.007}_{\pm 3.065}$	$20.853_{\pm 3.175}$
COCO512	$\mathbf{0.128}_{\pm 0.013}$	$0.133_{\pm 0.009}$	$\mathbf{32.893}_{\pm 0.457}$	$32.437_{\pm 0.229}$
GloVe25	$0.096_{\pm 0.021}$	$\mathbf{0.094}_{\pm 0.025}$	$9.737_{\pm 0.208}$	$\mathbf{9.830}_{\pm 0.115}$
GloVe100	$0.345_{\pm 0.094}$	$\mathbf{0.315}_{\pm 0.078}$	$\mathbf{8.310}_{\pm 0.242}$	$8.293_{\pm 0.283}$
LastFM65	$94.065_{\pm 0.098}$	$\mathbf{0.296}_{\pm 0.056}$	$1.977_{\pm 0.295}$	$\mathbf{17.687}_{\pm 0.721}$
NYTimes256	$0.221_{\pm 0.152}$	$\mathbf{0.219}_{\pm 0.144}$	$6.797_{\pm 0.307}$	$\mathbf{6.880}_{\pm 0.238}$
DBpedia1536	$\mathbf{0.202}_{\pm 0.038}$	$0.210_{\pm 0.042}$	$\mathbf{11.177}_{\pm 0.487}$	$\mathbf{11.177}_{\pm 0.392}$
COYO2048	$\mathbf{0.074}_{\pm 0.007}$	$0.075_{\pm 0.004}$	$12.813_{\pm 0.352}$	$\mathbf{12.913}_{\pm 0.374}$

For each condition, we alternate three timed repetitions of the two paths on the same GPU after warm-up. Timings start from CPU-resident inputs and include computing the origin, translation, encoder construction, transfers and full encoding. They exclude disk access, representation construction, diagnostics, bucket indexing and vector verification. We report mean and sample SD over three fits, using the within-fit median time. These measurements concern complete encoding, not end-to-end retrieval or join latency.

Recovering the specified nearest-reference rule. Across 729,600 diagnostic coordinates per path, common translation reduces disagreements with direct FP64 distances from 16,758 to five. The FP64 argmins of the translated FP32 diagnostic inputs agree with those of the original FP32 inputs throughout; the five remaining disagreements arise in score evaluation. These counts concern fixed diagnostic rows across related conditions, not independent observations or a database-wide error guarantee. All three repetitions produce identical full signatures within each path and condition.

The largest practical effect occurs on LastFM under the strongest compression (Table 16). Translation reduces the mean largest $b = 5$ bucket from 94.065% to 0.296% of the database and restores original-target retention from 1.977% to 17.687%. The fixed collision-count gate's retention rises from 82.500% to 95.823%. These recoveries follow from a more accurate evaluation of the same rule. Effects on load and retention are not uniformly monotone across the other datasets.

Native inputs expose both a benefit and a cost. Raw arithmetic has 169 diagnostic errors, all on LastFM; translation removes them but introduces one on DBpedia. Translation also increases complete encoding time on the two highest-dimensional datasets: the ratios of mean times are 1.845 on DBpedia and 2.062 on COYO. The numerical correction therefore requires a workload-dependent cost assessment; these encoding measurements do not revise the earlier retrieval timings.

Finally, numerical stability does not imply balanced buckets. With FixedSphere references at $\varepsilon = 0.01$, all 24 fits still have a $b = 5$ bucket containing the entire database after translation. That concentration is a property of the reference geometry. The intervention thus separates a correctable arithmetic failure from the geometric limits of data-dependent hashing.

E.6 Collision calibration and geometric locality

Calibration tests whether the collision formula predicts a fixed pair's random-index event. Locality additionally asks whether near pairs have larger agreement and whether accepted pairs are geometrically close. The latter depends on the pair distribution. We report full-pool calibration and certificate tightness together, followed by local probability evaluation; neither is a new retrieval experiment.

Table 17: **Large-pool collision calibration and geometric certificates have different accuracy.** Each dataset uses 16,384 references, 32 fixed pairs, all ten powers of two from $\psi = 2$ to 1024, and 512 independent $T = 100$ encoders per setting. Panel A averages absolute probability errors over pairs and settings, in percentage points; Max is the largest individual error across the three events. Panel B averages exact single-round probabilities and sufficient bounds over the 16 neighbor or 16 background pairs and all ten settings. This diagnostic compares quantities rather than competing methods, so no entries are ranked in bold. These averages do not define uniform locality constants or a corpus precision guarantee.

A: Probability error (percentage points)

Dataset	Single-round MAE	Band MAE	Count MAE	Max
Deep96	0.109	0.561	0.143	4.716
COCO512	0.106	0.514	0.087	5.206
LastFM65	0.101	0.548	0.150	4.280
NYTimes256	0.101	0.388	0.294	6.035

B: Exact probabilities versus geometric bounds

Dataset	Neighbor p_ψ	Margin lower	Background p_ψ	Coverage upper
Deep96	0.515	9.238×10^{-7}	0.115	1.000
COCO512	0.608	1.252×10^{-6}	0.119	0.998
LastFM65	0.334	0.000	0.083	1.000
NYTimes256	0.291	0.062	0.109	1.000

E.7 Calibration at the experimental reference-pool size

We extend the probability diagnostic from 64 to 16,384 references, using the actual reference pools of Deep96, COCO512, LastFM65 and NYTimes256. Each dataset supplies 16 fixed queries, one numerical top-10 neighbor per query with ranks cycled from one to ten, and one uniform database item outside that query's top ten. All 32 pairs share each fitted encoder. For every $\psi \in \{2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\}$, we draw 512 independent $T = 100$ encoders and evaluate single-round agreement, band admission with $b = 5$, and count admission with $m = 10$. These fixed diagnostic pairs use already observed queries; they do not constitute a new retrieval test or a corpus pair-prior estimate.

Predictions use Equation (17), with strict and weak two-coordinate rank counts evaluated by a grouped sweep. The sampled hashes are the first local minima of cached, subtraction-based FP64 squared distances. This numerical probability diagnostic is distinct from the production FP32 execution studies. Aggregate Monte Carlo uncertainty is computed across whole encoders, preserving dependence among pairs. Per-pair exact binomial tests use 51,200 trials for single-round agreement and 512 for each complete gate; Holm correction (Holm, 1979) covers all 3,840 tests at familywise level 0.05.

To assess geometric usefulness separately, write $r_{xy} = d(x, y)$ and evaluate the stronger endpoint margin bound

$$a_{xy} = \max_{u\in\{x,y\}} \mathbb{P}[\Delta_Z(u) > 2r_{xy}].$$

For the upper bound, let $\nu^\circ(u, s)$ be the probability that the sample misses the *open* ball of radius s around u, and use

$$c_{xy} = \min\{1, \nu^\circ(x, r_{xy}/2) + \nu^\circ(y, r_{xy}/2)\}.$$

If each endpoint has a sampled reference strictly inside its ball, a shared nearest reference would imply $d(x, y) < r_{xy}/2 + r_{xy}/2$, a contradiction. Thus $a_{xy} \le p_\psi(x, y) \le c_{xy}$ for each pair. These are pair-specific diagnostics, rather than uniform domain constants for Theorem 3.3 or a reverse-proximity certificate.

Table 17 separates prediction error from certificate tightness. Dataset-average errors are 0.101–0.109 percentage points for single-round agreement, 0.388–0.561 for band admission, and 0.087–0.294 for count admission. None of the 3,840 pointwise tests rejects after the Holm correction. The calibration therefore supports the finite-pool predictions at the reference-pool size used by the retrieval experiments.

The geometric certificates are substantially looser. On Deep, COCO and LastFM, the mean neighbor margin lower bound is below 2×10^{-6} despite mean exact agreement between 0.334 and 0.608. Mean background coverage upper bounds lie between 0.998 and 1.000 across the four datasets, while mean exact background agreement ranges from 0.083 to 0.119. The locality theorem supplies sufficient conditions to check; these diagnostics show that its margin and coverage bounds do not quantitatively explain most of the observed discrimination. Accurate collision calibration and useful geometric certification are separate requirements.

E.8 Small-pool Monte Carlo calibration

We calibrate single-round agreement and the complete band event on Sphere64, Mixture64, GloVe25, and COCO512. Exact finite-reference predictions are compared with 1,024 independently sampled encoders for each dataset and $\psi \in \{2, 8, 32\}$. The complete event uses $T = 100$ and $b = 5$. Dataset-average absolute errors are $(0.82\text{–}1.11) \times 10^{-3}$ for single-round agreement and $(2.54\text{–}5.81) \times 10^{-3}$ for candidate probability. These measurements test the direct law $1 - (1 - p_\psi^b)^{T/b}$ using the actual nearest-index signatures.

Appendix E.7 extends calibration to 16,384 references and $\psi \leq 1024$. Accurate prediction need not yield tight geometric bounds; Appendices A.7 and E.5 examine numerical selection and equivalent arithmetic.

E.9 Precision and local work at the full reference-pool size

We use all 32 fixed pairs from each of the four 16,384-reference pools in Appendix E.7: 16 query–neighbor pairs and 16 background pairs outside the query's top ten. We retain all ten ψ values from 2 to 1,024 and prescribed tolerances $\delta \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. For every ψ and tolerance, an exact binomial search finds the smallest k_0 satisfying $2q(k_0) \leq \delta$. Complete cutoff ties enter both neighborhoods. Local dominance counts determine S_I; rational probability endpoints retain the original population denominator.

An independent full-pool count and exact binomial recurrence verify the intervals. Small ordered-sample enumeration checks ties and boundary rankings. Candidate-gate endpoints are composed pairwise before averaging; the experiment introduces no new random encoders or retrieval fits.

For comparison, the geometric interval uses the pair-specific margin and coverage bounds from Appendix E.7. A count-only interval uses the two bounds preceding and following Equation (51). For each endpoint its candidate cutoff ranks are 1 and $\lceil an/\psi \rceil$, clipped to n, with $a \in \{1/4, 1/2, 1, 2, 4, 8, 16, 32\}$. The lower and upper cutoffs are optimized separately over their Cartesian grid. Thus this comparator may use different neighborhoods; the same-cutoff refinement result does not assert dominance over every separately optimized interval.

All 5,120 intervals contain the independently computed full-pool probability; all 1,280 primary intervals have single-round width at most 10^{-4}. Complete cutoff ties expand at least one neighborhood in 53 conditions; these expansions are included in all work measures. At the primary tolerance, the maximum union fractions for $\psi = 128, 256, 512, 1024$ are 14.844%, 7.532%, 3.772%, 1.868%, respectively. At $\psi = 2$, even the smallest observed union spans 99.292% of the pool.

The narrow local intervals reflect a prescribed tolerance, not an automatically stronger geometric theorem. Their work cost must be considered alongside their precision; the full-pool law remains the exact comparator. The background pairs are sampled outside the original top ten and need not be metrically far or semantically negative.

These intervals certify fixed-pair collision probabilities. Geometric separation, population bucket balance and new-query recall require additional evidence. Small ψ or cutoff ties can require nearly the whole pool. Figure 7 retains all four tolerances and ten reference-set sizes; the counting-domain reduction alone makes no full-sweep runtime claim.

F Complete computation and database growth

We separate complete retrieval on the fixed test queries from execution controls on development queries. The former compares deployed candidate pipelines; the latter identifies effects of resource

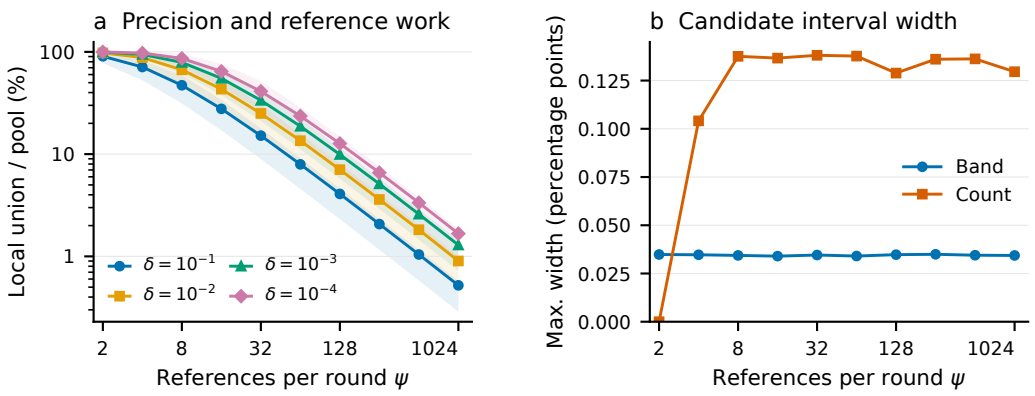

Figure 7: **Collision precision determines the required local neighborhood.** Left: union size $|U|/n$ at each tolerance, averaged over all 128 fixed pairs; shading spans their minimum and maximum, not a confidence interval. Right: maximum candidate-probability interval width across those pairs at $\delta = 10^{-4}$, with $T = 100$, $b = 5$, and count threshold $m = 10$. The universal bound is one percentage point. Neighborhood counting is local; the experiment still acquires distances to the full pool.

reuse, request size and intermediate storage. The growth studies keep fitted references and selection parameters fixed while expanding the database.

F.1 Construction, resource reuse, and candidate execution

The band-OR executor forms complete bucket unions with GPU bitsets, compacts candidates in database-ID order, and uses fused vector loads with FP64 reductions. Both hash families share the kernels. A paired control preserves every returned neighbor and candidate count across three SMinHash fits on each of the eight original datasets. This checks the fixed configurations, not a universal floating-point identity.

Table 7 includes per-index GPU initialization, sampling or training, and encoding or insertion in construction; common-vector transfer and selection are excluded. Flat and IVF-Flat query faster on all eight datasets. SMinHash has lower construction-plus-batch cost than IVF-Flat and HNSW in every fit at the displayed recalls. With shared GPU resources, Flat instead has lower total cost throughout; IVF-Flat wins on GloVe25, GloVe100 and LastFM, and SMinHash on the other five. Initialization policy therefore matters for persistent indexes.

Bitsets, population counts and prefix sums occupy $16\lceil N/32 \rceil$ bytes per query, beyond candidate/output arrays. Blocks target 64 MiB; oversized single queries remain intact. Candidates are never truncated, and word scans remain linear in N. Request-wise limits can allow different aggregate work at different batch sizes; none is triggered in this control.

F.2 Complete retrieval on the fixed test queries

This study measures complete retrieval for the eight 100k-item databases and 1,000 queries of Table 4, using requests of 32 and a final request of eight. Queries were already observed in the candidate-quality study; this is not a new test draw. Validation-selected parameters remain fixed for all three hash and IVF clustering fits.

Measurement and replication. Each fit compares SMinHash, Hyperplane1000, Flat and IVF-Flat in eager and CUDA Graph execution, with eight counterbalanced repetitions per arm. Timing includes copying resident queries, live encoding, complete candidate generation, scoring, top-10 selection, ID copying/concatenation and synchronization; build, warmup and capture preparation are excluded. Hash methods apply the fixed FP32 scorer to every candidate. Recall@10 uses returned IDs against

FP64 truth, so scoring precision can change recall even with the same candidates; Flat can also differ near ties.

Table 18 reports mean and sample SD of three within-fit median times. Flat's sessions replicate execution rather than fitting. This dispersion is not a confidence interval, and timing repetitions are not independent fits. Every planned configuration completed.

Devices and execution policy. Methods within each fit share an RTX 5880 Ada GPU and eight CPU threads. Fits span two or three physical GPUs on a shared host; dispersion therefore includes device and session effects, without exclusive-device assurances. Binary masks evaluate Hyperplane1000 and SMinHash at $\psi = 2$ or 16; other SMinHash settings use a complete packed-signature scan. All use the same fixed scoring and selection schedule. This policy changes execution only, preserving the frozen reference-index and count rules.

Correctness checks. Every signature and candidate count matches the candidate study. A separate CPU implementation checks complete candidates and FP32 top-10 scores for four queries per hash fit; every timing repetition's IDs match its untimed result. Both modes and all fits enter the aggregate.

F.3 Execution controls and workspace

A development-only control processes 200 reused queries on each of eight 100k databases at request sizes 1, 8, 32 and 200, comparing eager and CUDA Graph execution over five rotated passes of one fit. Timing includes resident-input copying, live encoding, complete candidate processing or native ANN search, and returned-ID copying. Build, warmup and capture preparation are excluded. Deep, DBpedia and COYO use selected execution schedules, whereas the other five use defaults; the tuning budgets differ.

Request size and capture change the ranking: SMinHash improves on Hyperplane1000 on LastFM, but Flat and IVF-Flat remain faster at larger requests. On DBpedia and COYO under capture, SMinHash beats Flat at sizes 32 and 200, while Hyperplane1000 is faster still. These development observations motivate the complete fixed-test comparison in Table 18; they are not additional test evidence.

Interval refinement uses guarded FP32 bounds with FP64 endpoints for finite FP32 inputs of dimension at most 4096. Unsafe cases use the original FP64 scorer; no candidate cutoff is imposed. The implementation reserves $40QN$ bytes for IDs, scores and bounds, plus gate/selection scratch. Sparse candidates reduce scoring work, not this allocation. Executable guards and complete development measurements accompany the code.

F.4 Intermediate storage and count execution

Fusing collision counting with its threshold writes a candidate bit mask without materializing a dense 16-bit count matrix, saving that $2QN$-byte intermediate. Compaction and verification retain their separate workspaces. The same optimization applies to both signature families and changes neither the nearest-index rule nor the complete candidate predicate.

A control on 200 reused development queries from NYTimes and GloVe100 uses 100k items, one fit and 15 interleaved full-query passes per arm. The separately selected signatures have $(T, \psi, m) = (1000, 2, 571)$ and $(250, 16, 52)$; these are not the primary $T = 100$ test configurations. For two references, the exact distance comparison is affine, $2x^\top(z_1 - z_0) \le \|z_1\|^2 - \|z_0\|^2$ for choosing local index zero, including equality. Equivalent arithmetic can differ near floating-point ties. Original and changed-query checks preserve the candidates, scores and returned IDs within each paired implementation control.

Fusion improves captured execution but has no uniform eager benefit. On NYTimes its original-encoder Graph median improves by 1.12×/1.30× at request sizes 32/200. With affine evaluation, Hyperplane1000 remains slightly faster at unrounded medians. On GloVe100 at size 200, fusion takes 0.968 ms versus Flat's 0.540 ms. These shared-device development measurements do not replace the three-fit fixed-test comparison or establish a general speed advantage.

Table 18: **Complete native retrieval with three fits.** Mean$_{\pm SD}$ of within-fit median times and returned recall on 1,000 test queries; eight timing repetitions per fit. Timing variability includes fitting, device, and session effects. The shaded rows identify SMinHash. Bold marks the highest recall; time minima are compared per dataset and mode among methods reaching 95% recall in all three fits. All planned arms completed; no individual repetitions are omitted.

Dataset	Method	Eager (ms)	Graph (ms)	Recall@10 (%)	Fits $\geq 95\%$
Deep96	SMinHash	$30.875_{\pm 1.602}$	$7.461_{\pm 0.543}$	$95.723_{\pm 0.704}$	3/3
	HP1000	$29.239_{\pm 0.481}$	$3.171_{\pm 0.084}$	$95.703_{\pm 0.097}$	3/3
	Flat	$6.943_{\pm 1.406}$	$6.354_{\pm 0.835}$	$\mathbf{100.000}_{\pm 0.000}$	3/3
	IVF-Flat	$\mathbf{5.399}_{\pm 0.299}$	$\mathbf{1.751}_{\pm 0.009}$	$96.910_{\pm 0.069}$	3/3
COCO512	SMinHash	$30.266_{\pm 3.617}$	$8.761_{\pm 0.081}$	$94.447_{\pm 0.470}$	0/3
	HP1000	$29.378_{\pm 0.572}$	$4.305_{\pm 0.325}$	$95.830_{\pm 0.894}$	2/3
	Flat	$14.869_{\pm 0.052}$	$14.615_{\pm 0.009}$	$\mathbf{100.000}_{\pm 0.000}$	3/3
	IVF-Flat	$\mathbf{5.039}_{\pm 0.144}$	$\mathbf{4.663}_{\pm 0.120}$	$97.303_{\pm 0.186}$	3/3
GloVe25	SMinHash	$29.419_{\pm 0.361}$	$7.111_{\pm 0.043}$	$95.517_{\pm 0.153}$	3/3
	HP1000	$28.616_{\pm 0.272}$	$2.910_{\pm 0.013}$	$94.150_{\pm 0.471}$	0/3
	Flat	$5.589_{\pm 0.009}$	$5.322_{\pm 0.018}$	$\mathbf{100.000}_{\pm 0.000}$	3/3
	IVF-Flat	$\mathbf{4.809}_{\pm 0.529}$	$\mathbf{1.628}_{\pm 0.031}$	$95.277_{\pm 0.140}$	3/3
GloVe100	SMinHash	$28.372_{\pm 0.819}$	$6.111_{\pm 0.065}$	$94.477_{\pm 0.301}$	0/3
	HP1000	$28.675_{\pm 0.647}$	$3.190_{\pm 0.019}$	$94.570_{\pm 0.922}$	1/3
	Flat	$6.272_{\pm 0.011}$	$5.952_{\pm 0.040}$	$\mathbf{100.000}_{\pm 0.000}$	3/3
	IVF-Flat	$\mathbf{5.330}_{\pm 0.026}$	$\mathbf{5.016}_{\pm 0.028}$	$98.027_{\pm 0.055}$	3/3
LastFM65	SMinHash	$28.461_{\pm 1.169}$	$6.977_{\pm 0.146}$	$95.823_{\pm 0.516}$	3/3
	HP1000	$30.140_{\pm 1.171}$	$28.137_{\pm 1.881}$	$95.263_{\pm 0.739}$	2/3
	Flat	$5.565_{\pm 0.017}$	$5.221_{\pm 0.012}$	$\mathbf{100.000}_{\pm 0.000}$	3/3
	IVF-Flat	$\mathbf{4.249}_{\pm 0.090}$	$\mathbf{2.099}_{\pm 0.036}$	$98.410_{\pm 0.121}$	3/3
NYTimes256	SMinHash	$31.875_{\pm 3.051}$	$9.347_{\pm 1.151}$	$95.367_{\pm 0.232}$	3/3
	HP1000	$29.326_{\pm 0.807}$	$\mathbf{3.058}_{\pm 0.088}$	$95.660_{\pm 0.305}$	3/3
	Flat	$\mathbf{9.437}_{\pm 0.153}$	$9.038_{\pm 0.206}$	$\mathbf{99.490}_{\pm 0.000}$	3/3
	IVF-Flat	$23.635_{\pm 0.650}$	$26.431_{\pm 0.919}$	$99.457_{\pm 0.035}$	3/3
DBpedia1536	SMinHash	$29.507_{\pm 2.859}$	$17.337_{\pm 3.589}$	$93.677_{\pm 1.035}$	0/3
	HP1000	$30.670_{\pm 3.167}$	$3.971_{\pm 0.097}$	$94.613_{\pm 0.576}$	1/3
	Flat	$\mathbf{33.615}_{\pm 3.485}$	$\mathbf{35.280}_{\pm 6.102}$	$\mathbf{100.000}_{\pm 0.000}$	3/3
	IVF-Flat	$46.230_{\pm 7.588}$	$45.298_{\pm 6.795}$	$94.673_{\pm 0.244}$	0/3
COYO2048	SMinHash	$28.293_{\pm 0.782}$	$18.328_{\pm 0.030}$	$94.510_{\pm 0.269}$	0/3
	HP1000	$30.827_{\pm 2.960}$	$4.752_{\pm 0.225}$	$94.590_{\pm 0.252}$	0/3
	Flat	$\mathbf{40.121}_{\pm 0.211}$	$\mathbf{40.376}_{\pm 0.215}$	$\mathbf{100.000}_{\pm 0.000}$	3/3
	IVF-Flat	$35.969_{\pm 0.074}$	$35.168_{\pm 0.139}$	$94.640_{\pm 0.190}$	0/3

F.5 Scaling a validation-selected candidate mechanism

Table 19 fixes the 100k validation selections, references, IVF training vectors and 1,000 query IDs while Deep grows to one million and 9.99 million items. SMinHash minimizes candidates and IVF-Flat minimizes query time at target 0.95. Top-10 truth is recomputed at each size. This experiment uses the original executor; its timings are separate from the paired execution control in Appendix F.1.

Candidate fraction stays near 3.10% while Recall@10 rises from 96.32% to 98.33% and 99.18%. At 9.99 million items, SMinHash takes 5.91 seconds to build and 7.45 seconds per batch, versus query times of 0.61 and 0.26 seconds for IVF-Flat and Flat. Fixed selectivity still entails more candidates as the database grows, and does not yield faster GPU search.

Table 19: **Fixed-setting scaling on Deep96.** Selections, reference/IVF training samples and 1,000 queries are fixed at 100k and reused at each scale. Recall@10 and candidates are percentages; build is in seconds and the complete batch in milliseconds. Values are mean$_{\pm SD}$ over completed fits, with coverage shown. — denotes uninstrumented candidate counts. Bold marks the best complete mean at each size, including displayed ties; achieved recall differs across methods.

N	Method	Recall (%)	Cand. (%)	Build (s)	Batch (ms)	Coverage
100k	SMinHash	$96.32_{\pm 0.12}$	$\mathbf{3.090}_{\pm 0.028}$	$\mathbf{0.04}_{\pm 0.00}$	$115.23_{\pm 0.72}$	3/3
	IVF-Flat	$98.21_{\pm 1.10}$	—	$0.36_{\pm 0.01}$	$\mathbf{3.03}_{\pm 0.23}$	3/3
	Flat (GPU)	$\mathbf{100.00}_{\pm 0.00}$	—	$0.15_{\pm 0.01}$	$3.74_{\pm 0.17}$	3/3
1M	SMinHash	$98.33_{\pm 0.02}$	$\mathbf{3.096}_{\pm 0.028}$	$0.64_{\pm 0.12}$	$688.69_{\pm 31.30}$	3/3
	IVF-Flat	$98.81_{\pm 0.78}$	—	$4.38_{\pm 2.08}$	$43.06_{\pm 3.34}$	3/3
	Flat (GPU)	$\mathbf{99.99}_{\pm 0.00}$	—	$\mathbf{0.29}_{\pm 0.03}$	$\mathbf{28.19}_{\pm 1.15}$	3/3
9.99M	SMinHash	$99.18_{\pm 0.11}$	$\mathbf{3.096}_{\pm 0.029}$	$\mathbf{5.91}_{\pm 0.11}$	$7451.21_{\pm 165.55}$	3/3
	IVF-Flat	$99.23_{\pm 0.57}$	—	$9.64_{\pm 3.92}$	$609.69_{\pm 14.50}$	3/3
	Flat (GPU)	$\mathbf{100.00}_{\pm 0.00}$	—	$10.52_{\pm 0.03}$	$\mathbf{261.44}_{\pm 9.92}$	3/3

F.6 Fixed-model transfer to larger embedding databases

We extend DBpedia1536 and COYO2048 from 100,000 to 990,000 entries, preserving the original database prefix and excluding all validation, test, and confirmation query IDs from the appended data. The original 16,384 reference vectors, three fitted models, validation-selected ψ and count thresholds, and the same 1,000 confirmation queries remain fixed. Only the additional 890,000 database vectors are encoded; the original database and query signatures are reused. We recompute numerical FP64 top-10 truth over all 990,000 vectors, using the same normalization, score tiles, and database-ID tie rule described in Appendix D.1. This is a paired transfer experiment on already evaluated queries, rather than another independent query confirmation or a search for new operating points.

Table 20: Candidate quality as the database grows with fixed models and queries. Values are mean$_{\pm SD}$ over three seeds. Bold marks the higher retention mean; candidate minima require 95% mean retention at that size. Displayed ties are included. Top-10 truth uses the full database at each size.

		Retained top-10 (%)		Candidates (%)	
Dataset	N	SMinHash	HP1000	SMinHash	HP1000
DBpedia1536	100k	$93.68_{\pm 1.04}$	$\mathbf{94.61}_{\pm 0.58}$	$0.758_{\pm 0.114}$	$0.196_{\pm 0.018}$
DBpedia1536	990k	$96.88_{\pm 0.50}$	$\mathbf{99.76}_{\pm 0.09}$	$0.758_{\pm 0.115}$	$\mathbf{0.195}_{\pm 0.018}$
COYO2048	100k	$94.51_{\pm 0.27}$	$\mathbf{94.59}_{\pm 0.25}$	$0.659_{\pm 0.001}$	$0.250_{\pm 0.027}$
COYO2048	990k	$97.69_{\pm 0.10}$	$\mathbf{99.59}_{\pm 0.06}$	$0.663_{\pm 0.001}$	$\mathbf{0.251}_{\pm 0.026}$

At 990k, SMinHash retains 96.88% and 97.69% of true neighbors on DBpedia and COYO, respectively, while admitting 0.758% and 0.663% of database items. All six SMinHash fits meet the original 95% target, compared with none at 100k. Hyperplane1000 also meets the target in

all six larger-database fits and remains more selective. The candidate fractions change little, but maintaining a similar fraction as the database grows still increases the absolute number of candidates by approximately the database-size ratio. Thus this result supports retention under fixed-model reuse; it does not establish a query-speed advantage, a size-independent verification cost, or linear-time all-pairs graph construction. Frozen parameters do not imply monotone retention: newly added nearest neighbors may select different references even when their metric distance to the query is smaller.