End-to-End Approximate Nearest Neighbour Search in a Distributed Transform-Domain Representation

Jonathan West[1]

[1]Independent Researcher, dtdr@multiverse1.com

Abstract

Approximate nearest neighbour (ANN) search is a core operation in modern vector databases, retrieval-augmented generation systems, and large-scale embedding pipelines. Most ANN methods operate on localised numerical representations, and performance in low-probe regimes is often limited by early information loss during coarse partitioning.

We investigate ANN search in a *distributed transform-domain representation* (DTDR), in which numerical vectors are represented using structured orthogonal transforms and quantisation, and treated as a primary computational domain rather than a transient compression format. We demonstrate an end-to-end ANN pipeline operating entirely in the DTDR domain, integrating inverted file indexing (IVF), per-partition HNSW traversal, and binary distance estimation, without reconstruction to full-precision vectors.

In this setting, we identify a multi-resolution signal, termed *dilution evidence*, arising from the persistence of similarity under progressive aggregation of distributed coefficients. This signal provides a coarse localisation cue that does not exist in conventional localised representations. Empirically, we show that dilution evidence recovers a substantial fraction of ANN recall in low-probe regimes, achieving recall comparable to higher-cost baselines while reducing query latency.

These results suggest that distributed transform-domain representations expose new design space for ANN search, enabling computational behaviour and optimisation signals that are unavailable in conventional vector representations.

1 Introduction

Approximate nearest neighbour (ANN) search underpins a wide range of modern systems [1] , including vector databases, retrieval-augmented generation (RAG) pipelines, recommendation engines, and large-scale semantic search. In these settings, queries are evaluated against collections containing millions or billions of high-dimensional vectors, requiring search strategies that trade exactness for tractable latency and memory usage.

Most contemporary ANN methods operate on *localised numerical representations*, in which information is concentrated in individual vector components or small groups of components. Indexing and traversal schemes such as inverted file (IVF) structures, graph-based methods (e.g. HNSW), and quantisation-based reranking exploit this locality to reduce the search space. However, performance in low-probe regimes is often limited by early information loss during coarse partitioning, where potentially relevant regions of the dataset are discarded before fine-grained similarity evaluation can occur.

This work explores ANN search in a *distributed transform-domain representation* (DTDR), in which numerical vectors are represented using structured orthogonal transforms and quantisation, and treated as a primary computational domain rather than a transient compression format. In DTDR, information is deliberately distributed across coefficients, such that no single component carries a disproportionate share of semantic content. This distributed structure gives rise to

different computational behaviour from conventional representations, particularly under partial aggregation or perturbation.

We demonstrate that ANN search can be performed end-to-end directly in the DTDR domain, without reconstruction to full-precision vectors. The proposed pipeline integrates inverted file indexing, per-partition HNSW traversal, and binary distance estimation, operating entirely on DTDR representations. This establishes DTDR not merely as a storage-efficient encoding, but as a viable computational substrate for ANN operations.

Within this setting, we observe an additional multi-resolution signal, which we term *dilution evidence*. This signal arises from the persistence of similarity under progressive aggregation of distributed coefficients, and provides a coarse localisation cue that is absent in conventional localised representations. Intuitively, similarity that survives aggregation is more likely to reflect globally relevant structure rather than coincidental local alignment.

Empirically, we show that dilution evidence recovers a substantial fraction of ANN recall in low-probe regimes, approaching the recall of higher-cost baselines while reducing query latency. These results suggest that distributed transform-domain representations expose new design space for ANN search, enabling optimisation signals and computational behaviour that are unavailable in conventional vector representations.

The full experimental code and reference implementations used in this work are available at `https://github.com/UnrealJon/DTDR`.

2 Background and Motivation

Approximate nearest neighbour (ANN) search addresses the problem of identifying vectors that are close to a query under a chosen similarity metric, typically cosine similarity or Euclidean distance, in high-dimensional spaces. Exact search scales poorly with dataset size, motivating a wide range of approximate methods that trade recall for reduced latency and memory consumption.

A common design pattern in ANN systems combines *coarse partitioning* with *fine-grained traversal or reranking*. Inverted file (IVF) methods partition the vector space into clusters [2] and restrict search to a subset of partitions, while graph-based approaches such as hierarchical navigable small-world (HNSW) graphs [3] enable efficient traversal within local neighbourhoods. Binary or quantised distance estimation is often used [4] to further reduce computational costs during candidate selection.

While these techniques are highly effective, their performance is sensitive to the number of partitions or regions probed during a query. In low-probe regimes, where only a small fraction of partitions are examined, relevant neighbours may be discarded early due to coarse assignment errors. Increasing the probe count improves recall but incurs higher latency and memory access costs, leading to a well-known recall–latency trade-off.

Most existing ANN methods implicitly assume *localised vector representations*, in which semantic information is concentrated in individual dimensions or small subsets of dimensions. Under this assumption, coarse partitioning decisions depend heavily on a limited number of components, and errors at this stage are difficult to recover. Subsequent traversal or reranking cannot compensate for information that has already been discarded.

By contrast, representations in which information is more globally distributed may exhibit different behaviour under coarse aggregation. If similarity is not dominated by a small number of components, partial or approximate evaluations may still retain useful information about global structure. This observation suggests the possibility of additional coarse localisation signals that are not available in conventional localised representations.

The work presented here is motivated by this hypothesis. Rather than modifying existing ANN algorithms, we investigate the consequences of performing ANN search directly in a distributed transform-domain representation. We ask whether treating such representations as a

primary computational domain can alter the recall–latency trade-off, particularly in low-probe regimes, and whether new signals emerge that can guide efficient search without increasing indexing complexity.

3 Distributed Transform-Domain Representation

Orthogonal transforms have long been used in high-dimensional data processing to redistribute information across coefficients while preserving inner products and norm structure . Such transforms provide well-understood numerical properties, including energy conservation and stability under quantisation, making them attractive as a computational substrate for large-scale vector operations. Similarity estimation in high-dimensional spaces is closely tied to inner product structure and randomized geometric transformations, which underpin many modern approximate nearest neighbour methods [5].

A distributed transform-domain representation (DTDR) encodes numerical vectors using structured orthogonal transforms followed by quantisation, with the result treated as a *primary stored and computational representation* rather than as a transient encoding. The defining characteristic of DTDR is that information is deliberately distributed across coefficients, such that semantic content is not concentrated in individual dimensions or small subsets of dimensions.

In contrast to localised representations, where individual components may dominate similarity computations, orthogonal transforms redistribute energy across the full coefficient set. As a result, no single coefficient carries a disproportionate share of information, and similarity arises from aggregate structure rather than isolated alignment. This distributed property persists under quantisation, provided the transform preserves orthogonality and global structure.

DTDR representations are constructed using fixed, structured orthogonal transforms, which may be applied uniformly across vectors of a given dimension. Quantisation is applied in the transform domain to reduce storage footprint and computational cost. Importantly, DTDR does not rely on entropy coding or variable- length representations; the transformed vectors retain a fixed dimensionality and support direct numerical operations.

A key consequence of this design is that DTDR representations remain suitable for computation without reconstruction to full-precision vectors. Inner products, distance estimates, and similarity measures can be approximated directly in the transform domain, enabling indexing, traversal, and reranking operations to be performed natively on DTDR vectors. Reconstruction to higher numerical precision, where required, can be performed as a separate and decoupled step.

From a computational perspective, DTDR representations exhibit behaviour that differs qualitatively from conventional encodings. Because information is globally distributed, partial aggregation, truncation, or perturbation of coefficients leads to smooth degradation of similarity estimates rather than catastrophic failure. This property underlies the robustness and graceful degradation observed in DTDR-based computations and motivates their use as a computational domain for approximate search.

In the context of ANN search, these properties suggest that DTDR may expose additional coarse-grained signals that are unavailable in localised representations. In particular, similarity that persists under progressive aggregation of distributed coefficients may provide information about global structure even when fine-grained detail is suppressed. The following sections explore how these properties can be exploited in an end-to-end ANN pipeline.

4 DTDR-Native ANN Pipeline

This section describes an end-to-end ANN search pipeline operating entirely in the distributed transform-domain representation. The objective is not to introduce new indexing structures,

but to demonstrate that established ANN components can be composed to operate natively on DTDR vectors without reconstruction to full-precision representations.

The pipeline follows a standard coarse-to-fine structure, consisting of (i) coarse partitioning using an inverted file (IVF) index, (ii) local traversal within selected partitions using graph-based search, and (iii) candidate reranking using low-cost distance estimation. Each stage operates directly on DTDR representations and preserves their distributed structure.

4.1 Coarse Partitioning via Inverted Files

Dataset vectors are assigned to a fixed number of coarse partitions using an IVF-style clustering scheme. Both dataset vectors and centroids are represented in the DTDR domain, and partition assignment is performed using approximate similarity evaluation in this domain. At query time, a subset of partitions is selected for further examination based on their coarse similarity to the query.

As in conventional IVF systems, the number of partitions probed (*nprobe*) controls the recall–latency trade-off. Low values of *nprobe* reduce latency but risk discarding relevant neighbours during coarse selection. The DTDR-native pipeline preserves this structure while enabling additional signals to be exploited at this stage, as discussed in Section 5.

4.2 Per-Partition Graph Traversal

Within each selected partition, local search is performed using a graph-based method analogous to hierarchical navigable small-world (HNSW) graphs. Separate graph indices are constructed for each partition using DTDR vectors as nodes. Queries traverse these graphs using approximate similarity evaluation in the DTDR domain, producing a set of candidate neighbours from each partition.

This per-partition traversal strategy limits graph size and traversal cost, while allowing fine-grained neighbourhood structure to be explored where coarse partitioning has identified potentially relevant regions. Importantly, all graph operations are performed without reconstructing vectors to higher numerical precision.

4.3 Binary Distance Estimation and Reranking

Candidate vectors returned from per-partition traversal are reranked using a low-cost distance estimation mechanism operating on compact DTDR-derived representations. Binary or quantised distance measures provide an efficient means of refining candidate sets prior to any optional higher-precision evaluation.

Because DTDR representations retain fixed dimensionality and support direct numerical operations, this reranking stage can be integrated seamlessly into the pipeline. Reconstruction to full-precision vectors, if required for downstream tasks, is decoupled from the search process and does not affect index structure or traversal logic.

4.4 End-to-End DTDR Operation

Across all stages, the DTDR representation is treated as the primary stored and computational form. No stage of the pipeline requires reconstruction to full-precision vectors in order to perform indexing, traversal, or candidate selection. This establishes DTDR as a viable computational domain for ANN search, rather than merely a storage-efficient encoding.

Crucially, operating entirely in the DTDR domain preserves the distributed structure of the representation throughout the search process. As a result, aggregation and approximation effects introduced during coarse partitioning or candidate selection exhibit smooth degradation rather than abrupt information loss. This behaviour underlies the additional localisation signal described in the following section.

5 Dilution Evidence

The distributed structure of DTDR representations gives rise to an additional coarse-grained signal that can be exploited during ANN search. We refer to this signal as *dilution evidence*, reflecting the fact that it emerges from the behaviour of similarity under progressive aggregation of distributed coefficients.

5.1 Intuition

In conventional localised vector representations, similarity is often dominated by a small subset of components. Coarse approximations that suppress or average over these components tend to destroy useful information abruptly, leading to early loss of recall when coarse partitioning decisions are incorrect.

By contrast, in DTDR representations, information is distributed across the full set of coefficients. As a result, partial aggregation or approximation does not eliminate similarity outright, but instead weakens it gradually. Similarity that persists under aggregation is therefore more likely to reflect globally relevant structure rather than coincidental alignment of a small number of components.

This behaviour suggests that similarity estimates evaluated at reduced resolution may still contain meaningful information about the location of nearest neighbours, even when fine-grained detail is suppressed. Dilution evidence exploits this observation by treating the persistence of similarity under aggregation as a coarse localisation cue.

5.2 Operational Definition

In the DTDR-native ANN pipeline, dilution evidence is computed by evaluating approximate similarity measures at multiple effective resolutions, obtained by progressively aggregating or subsampling transform-domain coefficients. At each resolution, similarity scores provide a noisy but inexpensive estimate of proximity between the query and candidate regions.

Regions or partitions that exhibit consistently elevated similarity across multiple resolutions are assigned higher confidence, while regions whose similarity collapses rapidly under aggregation are deprioritised. Importantly, this process does not introduce additional indexing structures or modify the underlying ANN components; it operates purely as an auxiliary signal derived from the DTDR representation itself.

The resulting dilution evidence signal is used to guide coarse partition selection and candidate prioritisation, particularly in low-probe regimes where standard IVF selection is most prone to early discard errors.

5.3 Relation to Existing ANN Signals

Dilution evidence differs fundamentally from conventional ANN optimisation signals such as centroid distance, graph traversal heuristics, or binary reranking scores. Those signals are typically evaluated at a single resolution and depend on localised component structure. Dilution evidence, by contrast, arises from multi-resolution behaviour intrinsic to distributed representations.

Crucially, this signal is not available in conventional vector encodings, where aggregation rapidly destroys discriminative information. It is therefore not an alternative parameterisation of existing ANN methods, but a consequence of treating DTDR as a primary computational domain.

5.4 Expected Behaviour

Because dilution evidence is derived from coarse aggregation, its effect is most pronounced in regimes where coarse selection dominates performance. In particular, the signal is expected to provide the greatest benefit at low values of *nprobe*, where conventional IVF-based methods suffer the largest recall degradation.

As the probe count increases and more partitions are examined, the marginal benefit of dilution evidence diminishes, since fine-grained traversal already recovers most relevant neighbours. This behaviour is consistent with the experimental results presented in the following section.

6 Experimental Evaluation

This section evaluates the behaviour of the DTDR-native ANN pipeline and the effect of dilution evidence on recall and latency, with particular emphasis on low-probe regimes. The experiments are designed to assess relative performance under controlled conditions rather than to optimise absolute performance on a specific benchmark.

6.1 Experimental Setup

We evaluate ANN search on a synthetic dataset of $N = 50{,}000$ vectors with dimensionality $d = 256$. Queries are drawn from the same distribution as the dataset vectors. Similarity is measured using cosine similarity.

All vectors are represented and stored in the distributed transform-domain representation. The ANN pipeline consists of an inverted file (IVF) index with $n_{\text{list}} = 256$ partitions, per-partition graph traversal using an HNSW-style index, and binary distance estimation for candidate reranking. Unless otherwise stated, HNSW parameters are fixed across experiments.

For each query, a fixed number of partitions (n_{probe}) are selected during coarse partitioning. We evaluate performance across a range of n_{probe} values to characterise the recall–latency trade-off. Dilution evidence, when enabled, is used solely as an auxiliary signal during coarse partition selection; all other components of the pipeline remain unchanged.

Recall is reported as recall@10, computed against exact nearest neighbours. Latency is reported as mean per-query wall-clock time.

6.2 Baseline Performance

As a baseline, we evaluate the DTDR-native ANN pipeline without dilution evidence. As expected, recall increases monotonically with n_{probe}, while query latency increases due to the larger number of partitions examined. In low-probe regimes, recall degradation is pronounced, reflecting early discard of relevant neighbours during coarse partitioning.

6.3 Effect of Dilution Evidence

Table 1 summarises the effect of dilution evidence at representative probe counts. At $n_{\text{probe}} = 4$, enabling dilution evidence improves recall@10 from 0.63 to 0.78, corresponding to an absolute gain of approximately 15 percentage points, while increasing mean query latency by approximately 0.2 ms.

For comparison, increasing n_{probe} from 4 to 8 without dilution evidence achieves a recall of 0.80 at a substantially higher latency. This demonstrates that dilution evidence recovers a large fraction of the recall otherwise obtained by probing additional partitions, at significantly lower cost.

Table 1: ANN search performance with and without dilution evidence.

Configuration	Recall@10	Latency (ms)
$n_{\text{probe}} = 2$, no dilution	0.50	2.36
$n_{\text{probe}} = 4$, no dilution	0.63	2.87
$n_{\text{probe}} = 4$, with dilution	**0.78**	3.10
$n_{\text{probe}} = 8$, no dilution	0.80	4.91

6.4 Observed Trends

The benefit of dilution evidence is most pronounced at low values of n_{probe}, where coarse partitioning decisions dominate recall. As n_{probe} increases, the marginal benefit diminishes, consistent with the expectation that fine-grained traversal recovers most relevant neighbours once a sufficient number of partitions are examined.

Importantly, the observed recall improvements are achieved without introducing additional index structures or increasing index complexity. Dilution evidence operates entirely as an auxiliary signal derived from the DTDR representation, and its computational cost remains modest relative to the overall query pipeline.

7 Discussion

The experimental results demonstrate that operating ANN search directly in a distributed transform-domain representation can expose additional optimisation signals beyond those available in conventional localised vector encodings. In particular, the dilution evidence signal provides a mechanism for recovering substantial recall in low-probe regimes, where coarse partitioning errors typically dominate performance.

A key observation is that dilution evidence does not replace existing ANN components, nor does it require modification of established indexing or traversal structures. Instead, it complements standard IVF and graph-based methods by providing an auxiliary coarse localisation cue derived from the behaviour of similarity under aggregation. This suggests that DTDR can be viewed as an enabling computational domain rather than as a competing ANN algorithm.

The effectiveness of dilution evidence in low-probe regimes is consistent with the intuition developed in earlier sections. When only a small number of partitions are examined, conventional coarse assignment decisions are brittle, and errors are difficult to recover. Because DTDR representations distribute information globally, approximate similarity evaluations retain meaningful signal even when evaluated at reduced resolution. The persistence of similarity under aggregation therefore provides information about global structure that is discarded in localised representations.

From a systems perspective, these results highlight an alternative axis of ANN optimisation. Rather than focusing exclusively on improved quantisation schemes, graph structures, or traversal heuristics, it may be advantageous to consider the choice of representation itself as a source of additional computational signals. In this view, DTDR does not merely reduce storage footprint, but alters the information geometry available to the search process.

It is also notable that the observed improvements are achieved with modest computational overhead. The additional latency incurred by dilution evidence is small relative to the overall query pipeline, while the recall gains are substantial in the regimes where they matter most. This trade-off suggests that DTDR-based signals may be particularly relevant in latency-sensitive applications, such as interactive retrieval and retrieval-augmented generation systems.

Finally, while the experiments presented here focus on a specific ANN pipeline, the underlying principles are more general. Any search or retrieval system that relies on coarse-to-fine evaluation may benefit from representations that preserve meaningful signal under aggregation. Distributed

transform-domain representations provide one concrete instantiation of this idea, but the broader implication is that representation choice can fundamentally influence the behaviour of approximate computation.

8 Limitations and Future Work

The results presented in this work are subject to several limitations that define the scope of the conclusions. First, the experimental evaluation is conducted on synthetic data, allowing controlled analysis of recall and latency behaviour but not capturing all characteristics of real-world embedding distributions. Evaluation on large-scale public benchmarks and production datasets remains an important direction for future work.

Second, the ANN pipeline examined here represents one specific composition of coarse partitioning, graph-based traversal, and candidate reranking. While this pipeline is representative of widely deployed ANN systems, alternative architectures or parameterisations may interact differently with distributed transform-domain representations. A systematic exploration of these design choices is beyond the scope of the present study.

Third, dilution evidence is evaluated as an auxiliary signal applied during coarse partition selection. Although the observed behaviour is consistent across the examined settings, the optimal manner in which this signal should be combined with existing ANN heuristics has not been fully explored. More sophisticated integration strategies, including adaptive or learned weighting, may yield additional gains.

Finally, this work focuses on the behaviour of DTDR representations during approximate search and does not address all aspects of end-to-end system integration. In particular, the interaction between DTDR-based ANN search and downstream tasks such as exact reranking, retrieval-augmented generation, or model inference warrants further investigation.

Future work will address these limitations by extending evaluation to real-world datasets, exploring alternative ANN architectures, and examining the interaction between distributed representations and other approximate computation tasks. These directions will help clarify the generality and practical applicability of the observed effects.

9 Conclusion

This work has examined approximate nearest neighbour search performed directly in a distributed transform-domain representation, treating the transform domain as a primary computational substrate rather than as a transient compression format. We have shown that an end-to-end ANN pipeline composed of established components can operate natively in this domain without reconstruction to full-precision vectors.

Within this setting, we identified an additional multi-resolution signal, termed dilution evidence, arising from the persistence of similarity under progressive aggregation of distributed coefficients. Empirical evaluation demonstrates that this signal recovers a substantial fraction of ANN recall in low-probe regimes, achieving performance comparable to higher-cost baselines while reducing query latency.

These findings suggest that distributed transform-domain representations expose new design space for approximate computation, enabling optimisation signals that are unavailable in conventional localised vector representations. Beyond ANN search, this perspective highlights the broader importance of representation choice in shaping the behaviour and efficiency of large-scale computational systems.

References

[1] R. Weber, H.-J. Schek, and S. Blott, "A quantitative analysis and performance study for similarity search methods in high-dimensional spaces," in *Proceedings of the 24th International Conference on Very Large Data Bases (VLDB)*, 1998, pp. 194–205.

[2] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, 2019.

[3] Y. A. Malkov and D. A. Yashunin, "Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[4] J. Li, Y. Sun, Z. Zhang *et al.*, "Rabitq: Quantization-based approximate nearest neighbor search with ranking refinement," *Proceedings of the VLDB Endowment*, 2023.

[5] M. S. Charikar, "Similarity estimation techniques from rounding algorithms," in *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, 2002, pp. 380–388.