MA-DPR: Manifold-aware Distance Metric for Dense Passage Retrieval

Anonymous ACL submission

Abstract

Dense Passage Retrieval (DPR) typically relies on Euclidean or Cosine distance to measure query-passage relevance in embedding space. While effective when embeddings lie on a linear manifold, our experiments across DPR benchmarks suggest that embeddings often lie on lower-dimensional, non-linear manifolds, especially in out-of-distribution (OOD) settings, where these distances fail to capture semantic similarity. To address this limitation, we propose a *manifold-aware* distance metric for DPR (**MA-DPR**) that models the intrinsic manifold structure of passages using a nearest neighbor graph and measures distance between query and passages based on their shortest path in this graph. We show that MA-DPR outperforms Euclidean and Cosine distance by up to 26% on OOD passage retrieval while maintaining performance on in-distribution data across various embedding models, with only a small increase in query inference time. Empirical evidence suggests that manifold-aware distance allows DPR to leverage context from related neighboring passages, making it effective even in the absence of direct semantic overlap. In addition, it can be extended to a wide range of dense embedding and DPR tasks, offering practical utility across diverse retrieval scenarios.

1 Introduction

Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) operates on the principle that semantically similar queries and passages remain close within a learned dense embedding space. By ranking passages based on their distances to the query, DPR measures the semantic relationships beyond direct word-level sparse matching. State-of-the-art DPR approaches primarily rely on Cosine distance and Euclidean distance (Mussmann and Ermon, 2016; Ram and Gray, 2012) due to their computational efficiency and straightforward interpretability.

However, the well-known manifold hypothesis states that high-dimensional data, such as text embeddings, reside on a subdimensional manifold that is often non-linear (Tenenbaum et al., 2000a; Roweis and Saul, 2000a). In such a case, relying solely on Euclidean and Cosine distance may fail to accurately capture the true relevance between queries and passages within this non-linear manifold structure (cf. Figure 1).

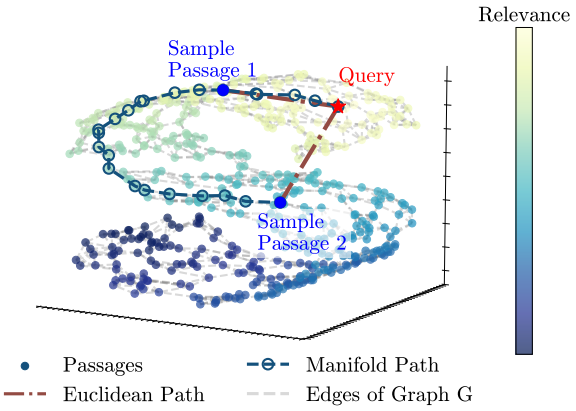

Figure 1: Example of subdimensional non-linear manifold in embedding space. Passages embedding (dots) form a non-linear S-shaped manifold, where their relevance (indicated by color) to the query (red star) is determined by proximity along the manifold rather than Euclidean distance. Two sample passages (blue dots) have similar Euclidean distance to the query (red path) but differ in relevance. In contrast, the distance (blue path) along the weighted undirected graph G—where nodes represent passages and edges (gray dashed lines) connect each passage to its K-nearest neighbors—better reflects the true relevance.

In contrast, a *manifold-aware* distance is intended to measure similarity along non-linear manifolds by leveraging graph-based (Zhou et al., 2003) or spectral (Belkin and Niyogi, 2003) methods that approximate the distances between points. While earlier work has acknowledged the presence of manifold and non-linear structures in embedding

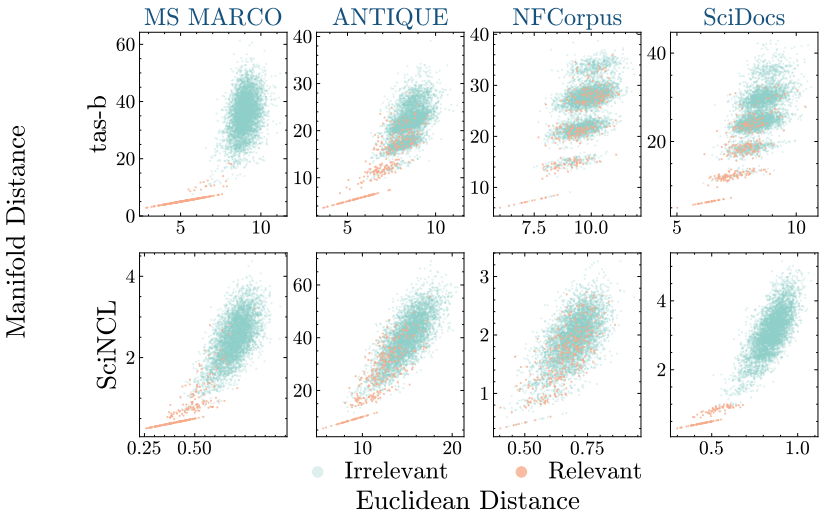

Figure 2: We evaluate the alignment between Euclidean distance (x-axis) and manifold-aware distance (y-axis) for all query-passage pairs in the embedding space across four benchmark datasets and two embedding models (`tas-b` and `SciNCL`), as detailed in Section 4. In-distribution datasets (MS MARCO for `tas-b` and SciDocs for `SciNCL`) exhibit strong alignment between the two distance metrics for relevant pairs (orange dots). In contrast, the three OOD datasets show clear misalignment, particularly at larger Euclidean distances, suggesting that $d_{\text{Euclidean}}$ fails to capture true semantic relationships in these settings.

spaces (Yonghe et al., 2019; Zhao et al., 2020), these important insights have been largely underexplored in the context of DPR. To the best of our knowledge, no prior work has applied manifold-aware distance metrics to DPR.

We hypothesize that a well-designed manifold-aware metric is better suited for capturing query-passage relevance in DPR where embeddings reside on a non-linear manifold. To this end, we summarize our contributions as follows:

1. We empirically verify the existence of sub-dimensional non-linear manifolds in out-of-distribution (OOD) embedding spaces, which significantly impacts the effectiveness of DPR when relying on standard Euclidean and Cosine distance (cf. Figure 2).

2. We introduce a generalizable manifold-aware distance metric for DPR (**MA-DPR**) that leverages a graph-based representation to exploit the non-linear manifold structure of the embedding space.

3. MA-DPR consistently outperforms DPR with Euclidean and Cosine distance across benchmark datasets and embedding models, with a comparable query inference time. By leveraging context from semantically related neighboring passages, it effectively retrieves relevant query-passage pairs even in the absence of direct semantic overlap.

2 Related Work

2.1 Dense Passage Retrieval

Dense Passage Retrieval (DPR) encodes queries and passages into a shared embedding space and ranks passages based on Euclidean and Cosine Distance (Mussmann and Ermon, 2016; Ram and Gray, 2012). These distance metrics are widely used due to their computational efficiency, interpretability, and compatibility with scalable approximate nearest neighbor search methods (Ram and Gray, 2012). The underlying assumption is that semantically relevant queries and passages are positioned close to each other in the embedding space.

However, this assumption is often violated, particularly in out-of-distribution (OOD) settings where embeddings are not optimized for the target domain. In such cases, Euclidean and Cosine distances may fail to capture complex semantic relationships, leading to retrieval failures despite close proximity in embedding space (Steck et al., 2024).

2.2 Manifold-Aware Approaches in Embedding Spaces

The manifold hypothesis posits that high-dimensional data, such as text embeddings,

reside on low-dimensional, non-linear manifolds embedded within the ambient space (Tenenbaum et al., 2000a; Roweis and Saul, 2000a). This makes manifold-aware distance metrics a more intrinsic solution for measuring similarity in such spaces compared to standard Euclidean and Cosine distances.

Previous studies, such as Isomap (Tenenbaum et al., 2000a), Locally Linear Embedding (LLE) (Roweis and Saul, 2000a), and Laplacian Eigenmaps (Belkin and Niyogi, 2003), approximate manifold-aware distances by constructing neighborhood graphs or computing spectral embeddings that capture the underlying manifold structure.

Similar manifold-aware approaches have been applied in Information Retrieval (IR) to incorporate global structural information, particularly in image retrieval tasks where the embedding space more naturally adheres to manifold structures (Zhou et al., 2003; He et al., 2004; Yang et al., 2013). However, manifold-aware distance metrics have not been systematically explored in DPR, leaving a critical yet unaddressed gap in the current retrieval framework.

3 MA-DPR: Manifold-Aware Distance for DPR

Let $\mathcal{P} = \{p^{(1)}, p^{(2)}, \ldots, p^{(N)}\}$ denote a collection of N passages, each associated with a dense embedding in a D-dimensional space, denoted as $\{\mathbf{e}_p^{(1)}, \ldots, \mathbf{e}_p^{(N)}\}$, where $\mathbf{e}_p^{(i)} \in \mathbb{R}^D$.

DPR ranks passages in $\mathcal{P}$ by measuring their semantic similarity to a given query q, computed as the distance between the query embedding $\mathbf{e}_q \in \mathbb{R}^D$ and each passage embedding $\mathbf{e}_p^{(i)}$ as:

$$\text{Rank}(q, \mathcal{P}) = \underset{p^{(i)} \in \mathcal{P}}{\text{argsort}}\ d(\mathbf{e}_q, \mathbf{e}_p^{(i)}), \qquad (1)$$

where $d(\cdot, \cdot)$ is a defined distance metric. In practice, common choices for d in DPR include Euclidean distance:

$$d_{\text{Euclidean}}(\mathbf{e}_q, \mathbf{e}_p^{(i)}) = \left\| \mathbf{e}_q - \mathbf{e}_p^{(i)} \right\|_2, \qquad (2)$$

and Cosine distance:

$$d_{\text{Cosine}}(\mathbf{e}_q, \mathbf{e}_p^{(i)}) = 1 - \frac{\mathbf{e}_q \cdot \mathbf{e}_p^{(i)}}{\|\mathbf{e}_q\|_2 \left\| \mathbf{e}_p^{(i)} \right\|_2}. \qquad (3)$$

Both $d_{\text{Euclidean}}$ and d_{Cosine} assume that semantic similarity corresponds to proximity in the embedding space, which may fail to capture the true semantic relationships between queries and passages in the presence of complex, non-linear structures.

Thus, we propose MA-DPR, an extension of DPR with a manifold-aware distance metric d_{Manifold}. MA-DPR operates in two stages: (1) a one-time *offline* manifold graph construction to capture the intrinsic manifold structure within embedding space, followed by (2) an *online* passage ranking stage for a given query based on the constructed graph.

3.1 Manifold Graph Construction

We propose a weighted undirected graph G to approximate the underlying non-linear manifold structure of the embedding space (cf. Figure 1) for $\mathcal{P}$. Each passage $p^{(i)} \in \mathcal{P}$ is represented as a vertex, and edges connect $p^{(i)}$ to its K-Nearest Neighbors (KNN) based on their proximity in the embedding space.

The use of KNN is motivated by its effectiveness in capturing local relationships while preserving the global structure of non-linear manifolds. This approach has been widely adopted in existing works of manifold structures (Tenenbaum et al., 2000b; Roweis and Saul, 2000b; Belkin and Niyogi, 2003) to construct graphs that reflect the local of high-dimensional data.

Specifically, let $G = (V, E, c)$, where:

- V: A set of vertices $\{v_1, \ldots, v_N\}$ representing $p^{(i)}, i \in \{1, \ldots, N\}$.
- E: A set of edges, where an edge $\{v_i, v_j\}$ exists between two vertices v_i and v_j if either $\mathbf{e}_p^{(i)}$ or $\mathbf{e}_p^{(j)}$ is among the KNN of the other based on a defined distance metric d^{KNN}.
- c: A cost function $c(v_i, v_j) : E \to \mathbb{R}$ assigns a cost to $\{v_i, v_j\}$.

Before introducing the passage ranking stage, we first present the proposed design choices for constructing the manifold graph: (1) the local distance metric d^{KNN} used to identify the K-nearest neighbors, (2) the edge cost function c, and (3) the number of nearest neighbors K.

3.2 Choices of Distance metrics d^{KNN}

Euclidean and Cosine distance When local neighbor distances are meaningful, as illustrated in Figure 1, the same distance metric used in DPR can be directly applied as d^{KNN}. Common choices

include the Euclidean distance $d^{\text{KNN}}_{\text{Euclidean}}$ (cf. Equation 2) and the Cosine distance $d^{\text{KNN}}_{\text{Cosine}}$ (cf. Equation 3).

However, as illustrated in the previous section, such distances in embedding space are not always reliable: $d^{\text{KNN}}_{\text{Euclidean}}$ and $d^{\text{KNN}}_{\text{Cosine}}$ may fail to capture the true underlying relationship between local neighbors. Moreover, it is sensitive to noise and outliers, as local perturbations can significantly affect the selection of K-nearest neighbors.

Spectral distance A widely adopted approach for d^{KNN} is to compute the distance between passages in a *spectral embedding space*, leveraging the *eigenstructure of the graph Laplacian* (Shi and Malik, 2000; Belkin and Niyogi, 2003).

Crucially, the spectral distance captures the intrinsic structure of the embedding manifold by leveraging global graph connectivity, rather than relying solely on local neighborhoods. This makes it well aligned with the manifold-aware objective of our work.

The *spectral distance* between two passages $p^{(i)}$ and $p^{(j)}$ is defined as:

$$d^{\text{KNN}}_{\text{Spectral}}\left(\mathbf{e}_p^{(i)}, \mathbf{e}_p^{(j)}\right) = \left\|\mathbf{u}^{(i)} - \mathbf{u}^{(j)}\right\|_2, \quad (4)$$

where $\mathbf{u}^{(i)} \in \mathbb{R}^m$ denotes the m-dimensional spectral embedding of passage $p^{(i)}$ obtained from the top m non-trivial eigenvectors of the normalized graph Laplacian.

By projecting the data into this lower-dimensional spectral space, the resulting distance metric can capture global structural relationships, enabling a more meaningful measure of between-node relationships, particularly in non-linear or noisy settings. This method has been widely applied for such purpose in NLP (Ploux and Ji, 2003; Dhillon et al., 2004; Tsai and Lee, 2016).

3.3 Choice of Cost Function c

Once the KNN Graph G is constructed as depicted in the graph mesh in Figure 1, we need to define the cost of edges E to compute distances during manifold-aware dense retrieval.

Distance Cost Distance Cost (DC) directly utilizes d^{KNN} as c:

$$c^{\text{DC}}(v_i, v_j) = d^{\text{KNN}}\left(\mathbf{e}_p^{(i)}, \mathbf{e}_p^{(j)}\right) \quad (5)$$

c^{DC} offers an embedding-aware distance for edge cost determination, but it is not clear that c^{DC} necessarily reflects ground truth differences in query relevance between two passages.

Uniform Cost Uniform Cost (UC) mitigates the influence of inaccurate ground truth differences from DC by adopting an agnostic (uninformed unit cost) perspective on the meaning of embedding distance. UC emphasizes discrete connectivity by counting the number of edges (hop) along the shortest path, focusing instead on mesh connectivity.

Specifically, all edges are assigned a constant scalar as cost to emphasize connectivity rather than actual distances as follows:

$$c^{\text{UC}}(v_j, v_j) = 1 \quad (6)$$

However, c^{UC} may discard useful information when neighbor distances accurately reflect differences in query relevance.

3.4 Choice of K

Selecting the value of K for KNN involves a trade-off: smaller values of K emphasize global structure by forming sparse graphs that reflect the broader topology of the manifold but may neglect local detail.

In contrast, larger values of K promote local connectivity by densely linking nearby nodes, which may oversimplify the manifold by treating it as locally linear.

Thus, Section 3.2 and Section 3.3 jointly define a two-fold design space for constructing the graph G. Section 4 presents empirical evaluations of each design choice (RQ3) and the impact of varying K on retrieval performance (RQ4) across a range of DPR benchmarks.

3.5 Passage Ranking

Given a constructed graph $G = (V, E, c)$, the manifold-aware distance d_{Manifold} between two passages is defined as the minimum total edge cost along any path connecting the corresponding vertices v_i and v_j in G (a.k.a. the shortest path).

d_{Manifold} is formally defined as follows:

$$d_{\text{Manifold}}(v_i, v_j) = \min_{\pi \in \Pi(v_i, v_j)} \sum_{(u,v)\in\pi} c(u, v), \quad (7)$$

where $\Pi(v_i, v_j)$ denotes the set of all paths from v_i to v_j, and $\pi \in \Pi(v_i, v_j)$ is a specific path represented as a sequence of edges $(u, v) \in E$.

To compute d_{Manifold} for a query q with embedding $\mathbf{e}_q$, q is temporarily added to G as a new vertex v_{N+1}. Edges are then formed by connecting

v_{N+1} to its K-nearest passage vertices using defined d^{KNN} based on $\mathbf{e}_q$.

For MA-DPR, each passage $p^{(i)}$ is ranked according to its manifold-aware distance from the query, $d_{\text{Manifold}}(v_{N+1}, v_i)$, A smaller value indicates a shorter path along the manifold in the embedding space, signifying higher relevance, while a larger $d_{\text{Manifold}}(v_{N+1}, v_i)$ implies lower relevance.

3.6 Computational Cost

MA-DPR introduces a one-time offline cost for constructing the manifold graph over the passage embeddings. This graph construction is independent of the query and does not affect the runtime efficiency of passage ranking at inference time.

At query time, the passage ranking process consists of two main steps: (1) computing distances between the query and all N passages to identify its K-nearest neighbors, with complexity $\mathcal{O}(ND)$; and (2) computing manifold-aware distances via shortest-path traversal on the KNN graph using Dijkstra's algorithm, which has complexity $\mathcal{O}(KN + N \log N)$, given $|E| = \mathcal{O}(KN)$.

Thus, the overall per-query complexity of manifold-aware distance DPR is: $\mathcal{O}(N(D + K + \log N))$.

In real-world retrieval settings where D ranges from 384 to 1024 and N reaches 10^5 passages, with typical K values between 2 and 15, the additional cost from graph traversal remains moderate. This makes manifold-aware distance a practical alternative for large-scale applications. A summary of per-query complexity and empirical latency is shown in Table 1.

Method	Complexity	Latency (100K)
$d_{\text{Euclidean}}$	$\mathcal{O}(ND)$	7.56 [7.56-7.56] ms
d_{Manifold}	$\mathcal{O}(N(D + K + \log N))$	8.10 [8.00–8.19] ms

Table 1: Per-query computational complexity and empirical latency of $d_{\text{Euclidean}}$ and d_{Manifold} (using $d^{\text{KNN}}_{\text{Euclidean}} + c^{\text{DC}}$, $K = 8$) for ranking over 100K passages with `tas-b` embeddings. Results include 95% confidence intervals in $[\cdot]$[1].

4 Experiments

4.1 Experimental Setup

Our experiments aim to evaluate the effectiveness of MA-DPR d_{Manifold} against baseline DPR with Euclidean distance ($d_{\text{Euclidean}}$)[2]. Specifically, we address the following key research questions:

RQ1: Nonlinear Manifold Validation Does empirical evidence support the presence of nonlinear manifold structure in dense embedding space?

RQ2: MA-DPR vs Bsaeline Does MA-DPR lead to improved retrieval performance compared to baseline DPR with $d_{\text{Euclidean}}$?

RQ3 Design Choice Comparison Which design choices for the manifold graph (cf. subsection 3.1) yield the best performance?

RQ4 Effect of K What is the effect of varying the number of K-nearest neighbors in the manifold graph on the performance of MA-DPRX?

RQ5 Underlying Factor Which type of queries does d_{Manifold} outperform $d_{\text{Euclidean}}$? What factors contribute to this improved performance?

We conduct experiments on four DPR benchmark datasets:

- MS MARCO (Nguyen et al., 2016)
- NFCorpus (Boteva et al., 2016)
- SciDocs (Cohan et al., 2020)
- ANTIQUE (Hashemi et al., 2020)

with two embedding models:

- `msmarco-distilbert-base-tas-b` (Hofstätter et al., 2021) (`tas-b`), trained on MS MARCO
- `SciNCL` (Ostendorff et al., 2022), trained on SciDocs.

The choice of embedding models naturally defines MS MARCO as the in-distribution dataset for `tas-b` and SciDocs as the in-distribution dataset for `SciNCL`, while the remaining datasets are treated as OOD. All embeddings are ℓ_2-normalized.[3]

For empirical evaluation, we assess the recall, Mean Average Precision (MAP), and Normalized Discounted Cumulative Gain (nDCG) for the top 20 ranked assignments of each retrieval result[4] .

[1]System specifications: CPU—Intel(R) Core(TM) i7-14700HX ; GPU—NVIDIA GeForce RTX 4070 Laptop GPU Average CPU utilization during measurement: ∼5%.

[2]With ℓ_2-normalized embeddings, Dot Product, Cosine, and Euclidean distance yield identical rankings. Thus, we report results using only one metric.

[3]Appendix A provides additional empirical analysis on the performance differences of the MA-DPR without normalization.

[4]To avoid the impact of parameter tuning, the number of

Table 2: Performance of d_{Manifold} across design choices on four datasets under fixed parameter settings ($k = 8$, $m = 700$). An underline indicates a statistically significant improvement of d_{Manifold} over $d_{\text{Euclidean}}$ (paired t-test, $p < 0.05$). We normalize `tas-b` and `SciNCL` embeddings so that $d_{\text{Euclidean}}$ and d_{Cosine} produce identical rankings.

	NFCorpus			**SciDocs**			**ANTIQUE**			**MS MARCO**		
	R@20	mAP@20	nDCG@20	R@20	mAP@20	nDCG@20	R@20	mAP@20	nDCG@20	R@20	mAP@20	nDCG@20
msmarco-distilbert-base-tas-b												
$d_{\text{Euclidean}}$ (d_{Cosine})	0.135	0.081	0.217	0.172	0.074	0.141	0.432	0.299	0.430	**0.950**	**0.534**	**0.638**
$d^{\text{KNN}}_{\text{Euclidean}}$+ c^{UC}	0.143	0.083	0.222	0.182	0.076	0.146	**0.467**	0.311	0.447	0.946	**0.534**	0.637
$d^{\text{KNN}}_{\text{Euclidean}}$+$c^{\text{DC}}$	0.137	0.083	0.220	0.167	0.074	0.139	0.417	0.299	0.424	0.945	**0.534**	0.637
$d^{\text{KNN}}_{\text{Spectral}}$+$c^{\text{UC}}$	**0.147**	**0.085**	0.223	**0.187**	**0.078**	**0.148**	0.464	**0.317**	**0.449**	0.944	**0.534**	0.636
$d^{\text{KNN}}_{\text{Spectral}}$+$c^{\text{DC}}$	**0.147**	**0.085**	**0.228**	0.182	0.077	0.146	0.436	0.307	0.430	0.932	0.498	0.604
SciNCL												
$d_{\text{Euclidean}}$ (d_{Cosine})	0.119	0.073	0.191	0.275	0.116	0.215	0.239	0.140	0.228	0.622	0.251	0.339
$d^{\text{KNN}}_{\text{Euclidean}}$+$c^{\text{UC}}$	**0.133**	**0.081**	0.203	0.266	0.116	0.211	**0.250**	**0.145**	**0.235**	**0.652**	**0.254**	**0.348**
$d^{\text{KNN}}_{\text{Euclidean}}$+$c^{\text{DC}}$	0.130	**0.081**	**0.205**	**0.279**	**0.119**	**0.217**	0.227	0.140	0.224	0.636	0.253	0.344
$d^{\text{KNN}}_{\text{Spectral}}$+$c^{\text{UC}}$	0.126	0.079	0.200	0.260	0.115	0.208	0.245	0.144	0.233	0.639	0.253	0.345
$d^{\text{KNN}}_{\text{Spectral}}$+$c^{\text{DC}}$	0.132	**0.081**	0.204	0.260	0.112	0.204	0.233	0.140	0.225	0.623	0.246	0.335

4.2 Experimental Results

All codes and results are available online[5].

RQ1 Nonlinear Manifold Validation: To empirically validate the non-linear manifold hypothesis in dense embedding spaces, we examine the relationship between d_{Manifold} and $d_{\text{Euclidean}}$ across relevant and irrelevant query-passage pairs. Specifically, in Figure 2, for each ground truth relevant query q and passage p pair (orange dots) and irrelevant pair (blue dots), we compute $d_{\text{Euclidean}}(q, p)$ and $d_{\text{Manifold}}(q, p)$ based on $d^{\text{KNN}}_{\text{Euclidean}}$ + c^{DC} for manifold graph construction.

In a perfectly linear embedding space, the manifold-aware distance induced by $d^{\text{KNN}}_{\text{Euclidean}}+c^{\text{DC}}$ should closely align with standard Euclidean distance. However, in the presence of non-linear structure, the two distances are expected to diverge. This contrast enables us to diagnose and characterize non-linear relationships in the embedding space.

Based on this intuition, we first observe a strong correlation between $d_{\text{Euclidean}}$ and d_{Manifold} in scatterplots of relevant query-passage pairs on the in-distribution datasets (i.e., MS MARCO w.r.t `tas-b` and SciDocs w.r.t. `SciNCL`). This correlation indicates that query-passage embeddings approximately lie on a locally linear manifold.

However, in the rest OOD datasets—where embeddings were not directly optimized during training—$d_{\text{Euclidean}}$ and d_{Manifold} exhibit significant misalignment, indicating that the embedding space follows a subdimensional non-linear manifold for both embedding models.

Further analysis reveals that in most OOD datasets, both relevant and irrelevant query-passage pairs fall within a similar range of $d_{\text{Euclidean}}$, causing them to be ranked similarly in DPR. This supports our conjecture (cf. Figure 1) that $d_{\text{Euclidean}}$ fails to capture the true semantic relationship between query and passage. In contrast, d_{Manifold} more effectively separates relevant and irrelevant query-passage pairs, with relevant pairs consistently exhibiting smaller d_{Manifold} values. These findings suggest that d_{Manifold} could better capture query-passage relationships by modeling the intrinsic non-linear manifold structure of the embedding space, motivating our further investigation in **RQ2**.

RQ2 MA-DPR vs Bsaeline: Motivated by **RQ1**, we empirically compare the performance of MA-DPR and DPR with $d_{\text{Euclidean}}$ and d_{Cosine}in Table 2.

Across nearly all four design choices, MA-DPR significantly outperforms baseline methods on OOD datasets. This aligns with RQ1, where OOD datasets exhibit a subdimensional non-linear manifold structure. In such cases, d_{Manifold} more effectively captures the underlying data manifold structure and appears to better capture relevance.

In contrast, on in-distribution datasets—where the embedding space approximates a locally linear manifold, as empirically shown in RQ1—both d_{Manifold} and $d_{\text{Euclidean}}$ yield similar retrieval performance. In such cases, d_{Manifold} effectively reduces

neighbors K in the KNN graph is fixed to 8 for all experiments except RQ4, and the spectral embedding dimension M is fixed to 700 throughout.

[5] anonymous.4open.science/r/Manifold_distance_Retrieval-F226

to $d_{\text{Euclidean}}$, as the shortest path along the manifold aligns with the Euclidean distance (cf. Figure 2).

Thus, MA-DPR achieves superior performance on OOD datasets without parameter tuning, while remaining competitive in in-distribution settings, offering a more robust and generalizable alternative to Euclidean and Cosine distance without additional tuning.

In addition, MA-DPR outperforms DPR with Euclidean and Cosine distance across both embeddings, indicating that its effectiveness is not tied to a specific embedding space. This improvement underscores its robustness to variations in the embedding model, further validating its generalizability.

RQ3 Design Choice Comparison: Table 2 also empirically evaluates the performance of different design choices within d_{Manifold} as discussed in Section 3.1. Results indicate that performance is variable across datasets and embedding models.

For most OOD datasets, c^{UC} outperforms c^{DC}, as it prioritizes connectivity over raw distances in embedding space, which are often unreliable due to distortions in out-of-distribution settings. c^{UC} emphasizes the discrete transitions between neighboring passages rather than relying on possibly misleading distances. This is particularly beneficial when relevant passages are not directly similar to the query but lie along a chain of semantically related neighbors, which will be further discussed in RQ5 (cf. Figure 4).

The performance between $d^{\text{KNN}}_{\text{Euclidean}}$ and $d^{\text{KNN}}_{\text{Spectral}}$ varies across embedding models. $d^{\text{KNN}}_{\text{Euclidean}}$ performs better with `SciNCL` since `SciNCL` explicitly preserves local neighborhood structure through neighbor-aware contrastive learning (Ostendorff et al., 2022) as its training objective, making $d^{\text{KNN}}_{\text{Euclidean}}$ a strong fit for capturing local similarity.

In contrast, `tas-b` benefits more from $d^{\text{KNN}}_{\text{Spectral}}$ since it is optimized for global ranking (Hofstätter et al., 2021), which captures manifold-wide structure. These results support our hypothesis (cf. subsection 3.2) that different graph construction metrics capture complementary properties of dense embedding spaces: $d^{\text{KNN}}_{\text{Euclidean}}$ is better suited for locally organized structures, whereas $d^{\text{KNN}}_{\text{Spectral}}$ is more effective in capturing globally structured manifolds.

RQ4 Effect of K: #K neighbors controls the balance between preserving local manifold structure and maintaining global connectivity of the graph (cf. subsection 3.1), which can influence the effectiveness of the manifold-aware distance and lead to performance variations. Figure 3 evaluate the performance of MA-DPR with $K \in \{1, \ldots, 15\}$(in nDCG only, see Appendix B for full results).

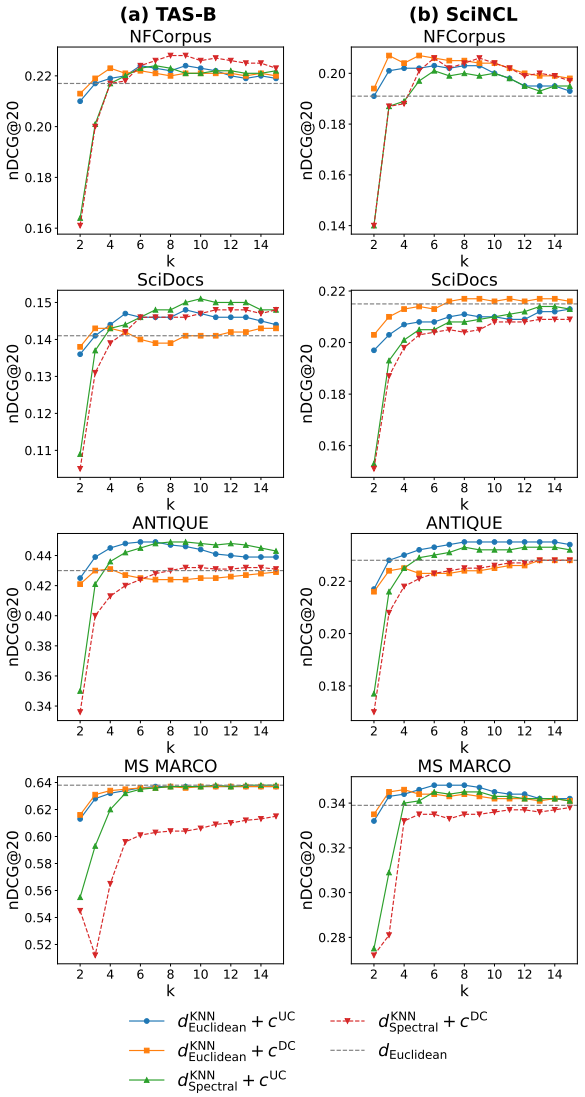

Figure 3: Performance of MA-DPR in nDCG@20 across varying values of #K neighbors (x-axis).

At low K, the graph captures only very local relationships, which may be too sparse to support effective global retrieval. At high K, the graph becomes overly dense, and shortest-path distances begin to approximate Euclidean distance. Intermediate values of K strike a favorable balance, preserving local structure while maintaining sufficient connectivity for manifold-aware traversal.

RQ5 Underlying Factor: Figure 4 presents the distribution of d_{Manifold} and $d_{\text{Euclidean}}$ across their respective top 500 retrieved passages for two exam-

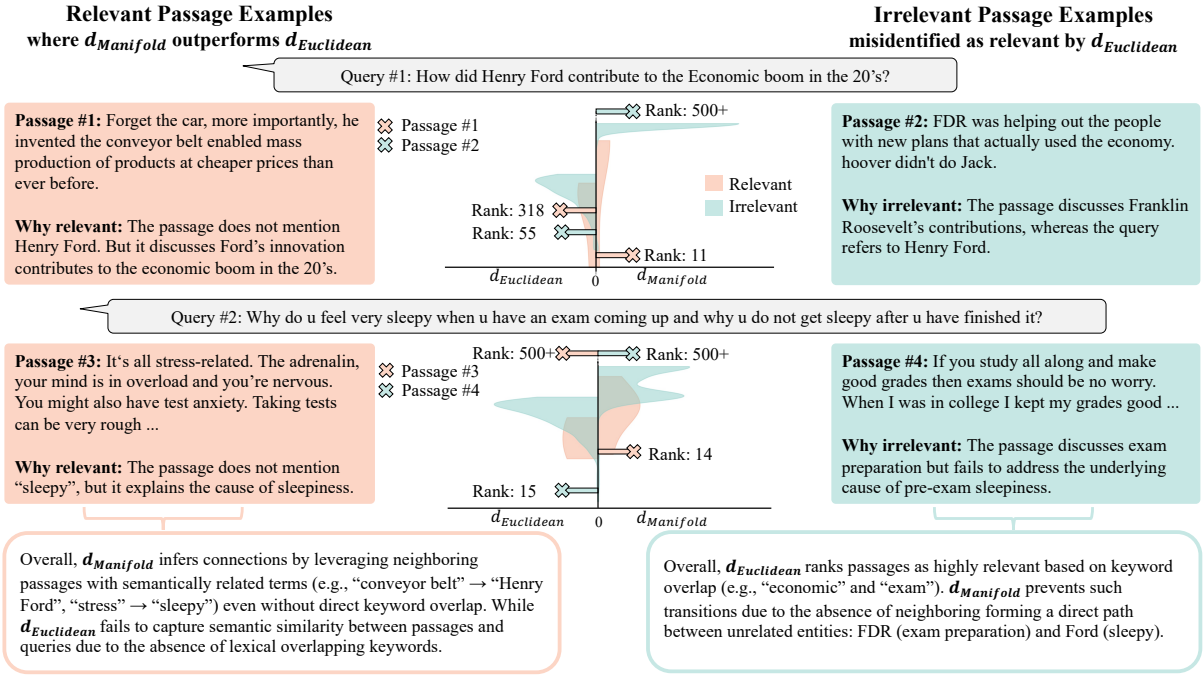

Figure 4: Examples from ANTIQUE where MA-DPR outperforms DPR with $d_{\text{Euclidean}}$ under `tas-b`. We present (i) relevant passages successfully retrieved by d_{Manifold} within the top 20 but missed by $d_{\text{Euclidean}}$, and (ii) irrelevant passages that misidentified as relevant by $d_{\text{Euclidean}}$. The Kernel Density illustrates the distribution of $d_{\text{Euclidean}}$ and d_{Manifold} distances among the top 500 retrieved passages, categorized by ground truth relevance. $d_{\text{Euclidean}}$ exhibits substantial overlap between relevant and irrelevant passages, failing to distinguish true relevance. In contrast, d_{Manifold} demonstrates clear separation.

ple queries. In both cases, $d_{\text{Euclidean}}$ fails to clearly distinguish relevant passages from irrelevant ones in the density plots (middle), as their distance distributions significantly overlap. In contrast, d_{Manifold} effectively separates relevant passages into a distinct range.

Further analysis of the context in these queries and passages (cf. text in Figure 4) reveals cases where $d_{\text{Euclidean}}$ has poor performance: (1) settings where relevant Passages 1 and 3 require reasoning and lack direct semantic overlap with the query, which $d_{\text{Euclidean}}$ fails to identify as relevant; and (2) settings such as irrelevant Passage 4 that contains partially overlapping keywords but different contextual meanings, or Passage 2 where two historical figures have a strong economic association. Such misleading similarities cause $d_{\text{Euclidean}}$ to erroneously rank them as relevant.

In contrast, d_{Manifold} remains effective in all cases of Figure 4: (1) For relevant Passages 1 and 3, d_{Manifold} leverages neighboring passages to provide crucial missing context that bridges the query and passage in the absence of direct similarity. (2) For irrelevant Passages 2 and 4 with misleading lexical or semantic overlap, the lack of semantically similar neighboring passages precludes d_{Manifold} from small distances, and hence relevance to the query.

5 Conclusion

With the aim to better capturing relevance in DPR via distance on the subdimensional non-linear manifold of query and passage embeddings, we introduced a novel distance metric to leverage the underlying manifold structure of embeddings using a graph-based representation.

With a one-time computational cost for graph construction and comparable online query inference cost to standard DPR, our proposed MA-DPR is able to exploit the manifold structure of embedding space and achieves a 26% improvement over DPR using traditional Euclidean and cosine distances, particularly on OOD datasets.

By leveraging the context from neighboring passages, manifold-aware DPR demonstrates effectiveness for queries and passages lacking overlapping keywords. These findings suggest that manifold-aware distance can significantly enhance DPR performance and that search over the implicit manifold of data can help overcome deficiencies in embedding training for OOD settings.

6 Limitations

While the proposed MA-DPR effectively models non-linear structures in embedding space, several limitations remain. First, the current approach relies on a flat KNN graph constructed offline, which may pose scalability challenges as the number of passages grows. Future work can address this by exploring hierarchical or multi-resolution graph structures to reduce traversal cost while preserving manifold properties. Second, the graph edge weights are derived from unsupervised distance metrics, which may not always align with relevance signals in retrieval tasks. Incorporating supervised signals to learn edge weights can further refine the manifold representation and improve retrieval effectiveness.

References

Mikhail Belkin and Partha Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396.

Viktoriya Boteva, Ralf Schenkel, and Avishek Anand. 2016. A full-text learning to rank dataset for medical information retrieval. *Proc. of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Inderjit S. Dhillon, Guan Yuqiang, and Brian Kulis. 2004. A unified view of graph-based semi-supervised learning and spectral clustering. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*.

Hamed Hashemi, W Bruce Croft, and David Jensen. 2020. Antique: A non-factoid question answering benchmark. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM)*.

Jingrui He, Mingjing Li, Hong-Jiang Zhang, Hanghang Tong, and Changshui Zhang. 2004. Manifold-ranking based image retrieval. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 9–16.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling. In *Proc. of SIGIR*.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Stephen Mussmann and Stefano Ermon. 2016. Learning and inference via maximum inner product search. In *International Conference on Machine Learning*, pages 2587–2596. PMLR.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human-generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. *Preprint*, arXiv:2202.06671.

Sabine Ploux and Hong Ji. 2003. Semantic spaces: Modeling the distribution of word senses with unsupervised learning techniques. *Computational Linguistics*, 29(2):259–275.

Parikshit Ram and Alexander G. Gray. 2012. Maximum inner-product search using cone trees. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 931–939, New York, NY, USA. Association for Computing Machinery.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sam T. Roweis and Lawrence K. Saul. 2000a. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326.

Sam T Roweis and Lawrence K Saul. 2000b. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326.

Jianbo Shi and Jitendra Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM on Web Conference 2024*, pages 887–890.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. 2000a. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. 2000b. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323.

Cheng-Lung Tsai and Yu-Chiang Frank Lee. 2016. Multinomial pca for learning topic-based document semantics. In *Proceedings of ACL*, pages 1525–1534.

Cheng Yang, Li Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2013. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3166–3173. IEEE.

Chu Yonghe, Hongfei Lin, Liang Yang, Yufeng Diao, Shaowu Zhang, and Fan Xiaochao. 2019. Refining word reesprentations by manifold learning. In *Proc. 28th Int. Joint Conf. Artif. Intell*, pages 5394–5400.

Wenyu Zhao, Dong Zhou, Lin Li, and Jinjun Chen. 2020. Manifold learning-based word representation refinement incorporating global and local information. In *Proceedings of the 28th international conference on computational linguistics*, pages 3401–3412.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. 2003. Ranking on data manifolds. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 16, pages 169–176. MIT Press.

A Impact of Normalization

In this section, we examine the effect of embedding normalization on the performance of our proposed MA-DPR.

As shown in Table 3, we observe that removing normalization leads to a performance drop for both the baseline $d_{\text{Euclidean}}$ and d_{Manifold}, with the degradation occurring at a similar level. Importantly, the performance gap between d_{Manifold} and the baseline remains consistent, suggesting that MA-DPR is robust to whether embeddings are normalized or not.

Table 3: Performance of d_{Manifold} across design choices on four datasets under fixed parameter settings ($k = 8$, $m = 700$). An underline indicates a statistically significant improvement of d_{Manifold} over $d_{\text{Euclidean}}$ (paired t-test, $p < 0.05$). We use original embeddings without normalization.

	NFCorpus			**SciDocs**			**ANTIQUE**			**MS MARCO**		
	R@20	mAP@20	nDCG@20	R@20	mAP@20	nDCG@20	R@20	mAP@20	nDCG@20	R@20	mAP@20	nDCG@20
msmarco-distilbert-base-tas-b												
$d_{\text{Euclidean}}$	0.110	0.065	0.185	0.155	0.065	0.127	0.409	0.283	0.412	**0.945**	**0.526**	**0.630**
$d^{\text{KNN}}_{\text{Euclidean}}$+ c^{UC}	0.123	0.070	0.196	0.166	0.068	0.131	**0.451**	0.298	0.431	0.944	**0.526**	**0.630**
$d^{\text{KNN}}_{\text{Euclidean}}$+$c^{\text{DC}}$	0.111	0.069	0.193	0.148	0.065	0.124	0.389	0.282	0.402	0.940	0.525	0.629
$d^{\text{KNN}}_{\text{Spectral}}$+$c^{\text{UC}}$	**0.129**	0.072	**0.199**	**0.175**	**0.070**	**0.136**	0.448	**0.303**	**0.434**	0.944	**0.526**	0.629
$d^{\text{KNN}}_{\text{Spectral}}$+$c^{\text{DC}}$	0.126	**0.074**	0.197	0.159	0.068	0.126	0.385	0.279	0.387	0.928	0.483	0.601
SciNCL												
$d_{\text{Euclidean}}$	0.120	0.073	0.190	0.279	0.117	0.217	0.238	0.138	0.227	0.616	0.250	0.338
$d^{\text{KNN}}_{\text{Euclidean}}$+$c^{\text{UC}}$	**0.131**	**0.080**	0.202	0.269	0.117	0.212	**0.249**	**0.144**	**0.233**	**0.652**	**0.254**	**0.348**
$d^{\text{KNN}}_{\text{Euclidean}}$+$c^{\text{DC}}$	0.130	**0.080**	**0.204**	**0.281**	**0.119**	**0.218**	0.224	0.138	0.221	0.630	0.252	0.342
$d^{\text{KNN}}_{\text{Spectral}}$+$c^{\text{UC}}$	**0.131**	0.079	0.201	0.260	0.116	0.209	0.243	0.143	0.231	0.636	0.253	0.344
$d^{\text{KNN}}_{\text{Spectral}}$+$c^{\text{DC}}$	0.117	**0.080**	0.196	0.226	0.099	0.180	0.203	0.126	0.200	0.577	0.229	0.311

B Impact of hyperparameter

We present the full results of the effect of K on $d^{\text{KNN}}_{\text{Euclidean}}$ using mAP and Recall in Figure 5 and Figure 6, which show a similar trend to the nDCG results discussed in RQ4.

We also report the effect of spectral dimension M on $d^{\text{KNN}}_{\text{Spectral}}$ in Figure 7 and Figure 8.

C Additional Embedding

We additionally report results using `msmarco-distilbert-dot-v5` (Reimers and Gurevych, 2019) to provide further empirical support for our method, trained on MS MARCO. The results are consistent with our main findings and align with the discussions presented in the paper (cf. Table 4).

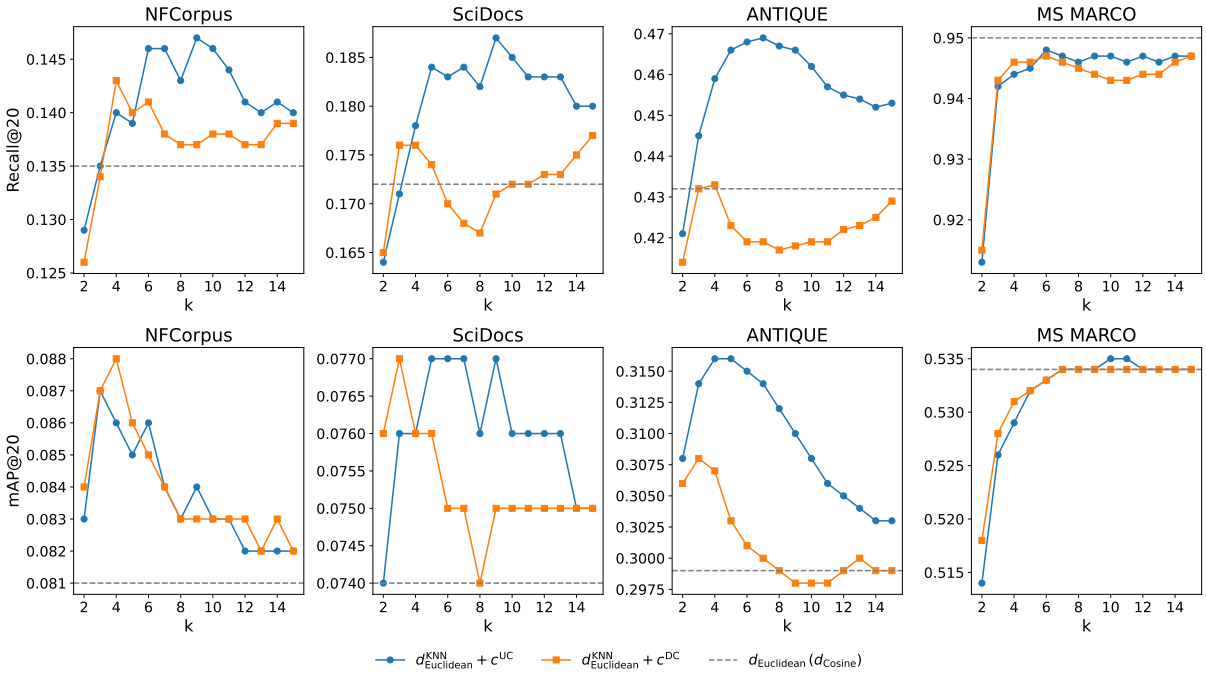

Figure 5: Recall@20 and mAP@20 comparison of $d^{\text{KNN}}_{\text{Euclidean}}$ with TAS-B across varying k.

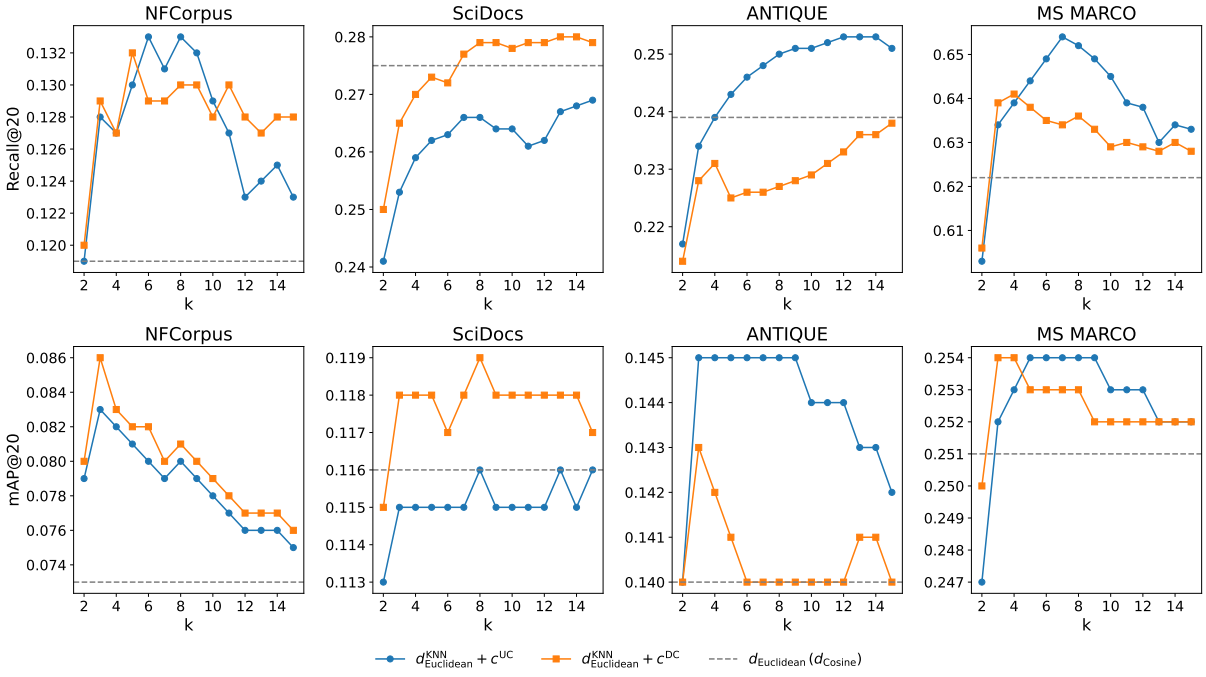

Figure 6: Recall@20 and mAP@20 comparison of $d^{\text{KNN}}_{\text{Euclidean}}$ with SciNCL across varying k.

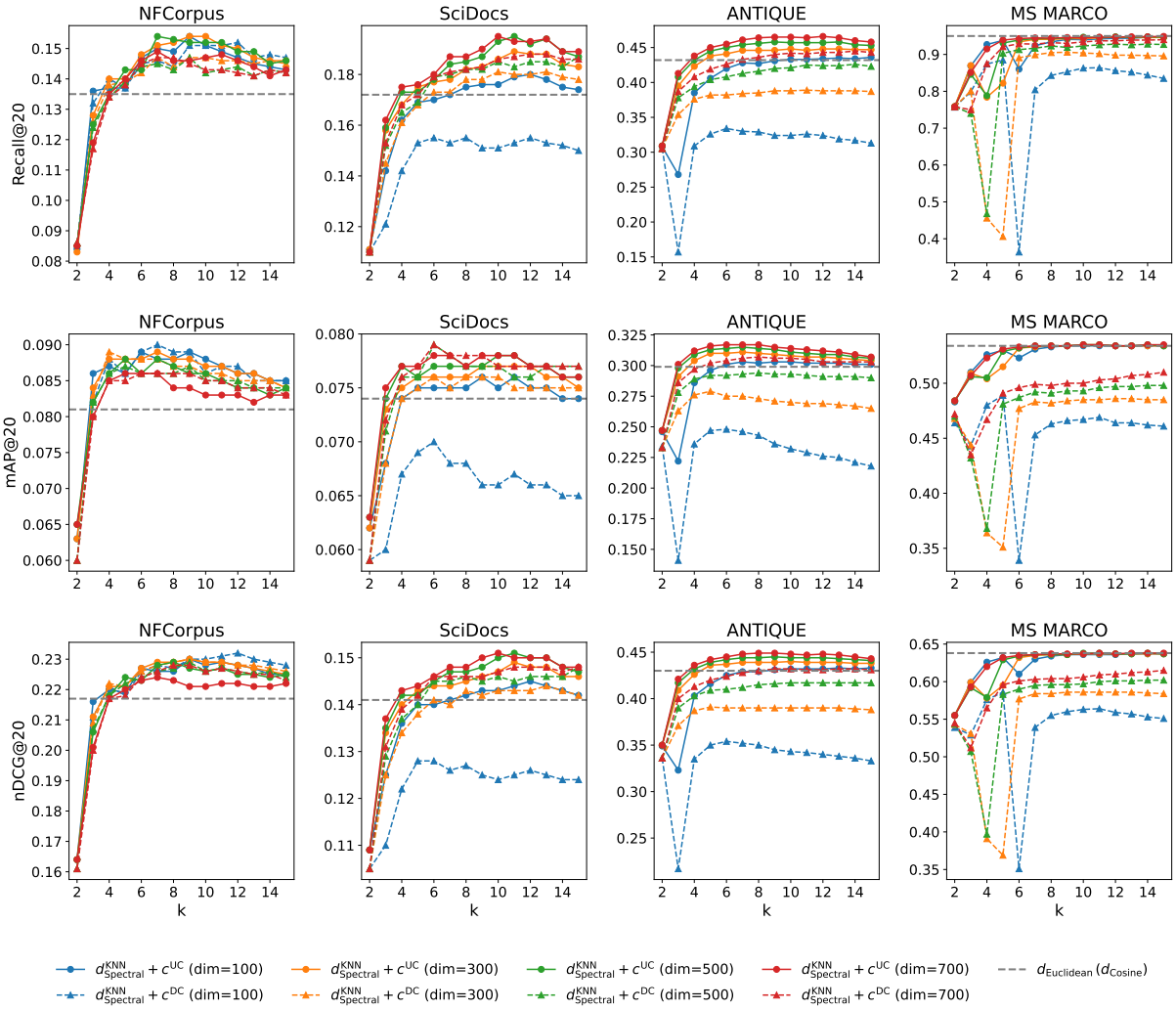

Figure 7: Recall@20, mAP@20 and nDCG@20 comparison of $d^{\text{KNN}}_{\text{Spectral}}$ with TAS-B across varying k.

Table 4: Performance of d_{Manifold} across design choices on four datasets under fixed parameter settings ($k = 8$, $m = 700$). An underline indicates a statistically significant improvement of d_{Manifold} over $d_{\text{Euclidean}}$ (paired t-test, $p < 0.05$). We normalize `dot-v5` embeddings so that $d_{\text{Euclidean}}$ and d_{Cosine} produce identical rankings.

	NFCorpus			**SciDocs**			**ANTIQUE**			**MS MARCO**		
	R@20	mAP@20	nDCG@20	R@20	mAP@20	nDCG@20	R@20	mAP@20	nDCG@20	R@20	mAP@20	nDCG@20
msmarco-bert-base-dot-v5												
$d_{\text{Euclidean}}$ (d_{Cosine})	0.132	0.071	0.192	0.168	0.070	0.138	0.382	0.251	0.377	0.937	**0.529**	0.631
$d^{\text{KNN}}_{\text{Euclidean}} + c^{\text{UC}}$	0.133	0.077	0.202	0.181	0.079	0.147	0.423	0.268	0.398	0.942	**0.529**	0.632
$d^{\text{KNN}}_{\text{Euclidean}} + c^{\text{DC}}$	0.133	0.078	0.204	0.181	0.079	0.147	0.365	0.253	0.372	0.936	**0.529**	0.631
$d^{\text{KNN}}_{\text{Spectral}} + c^{\text{UC}}$	0.136	0.077	0.203	**0.183**	**0.080**	**0.148**	**0.427**	0.275	**0.404**	**0.943**	**0.529**	**0.633**
$d^{\text{KNN}}_{\text{Spectral}} + c^{\text{DC}}$	**0.139**	**0.078**	**0.207**	0.182	**0.080**	**0.148**	0.411	**0.277**	0.399	0.929	0.499	0.604

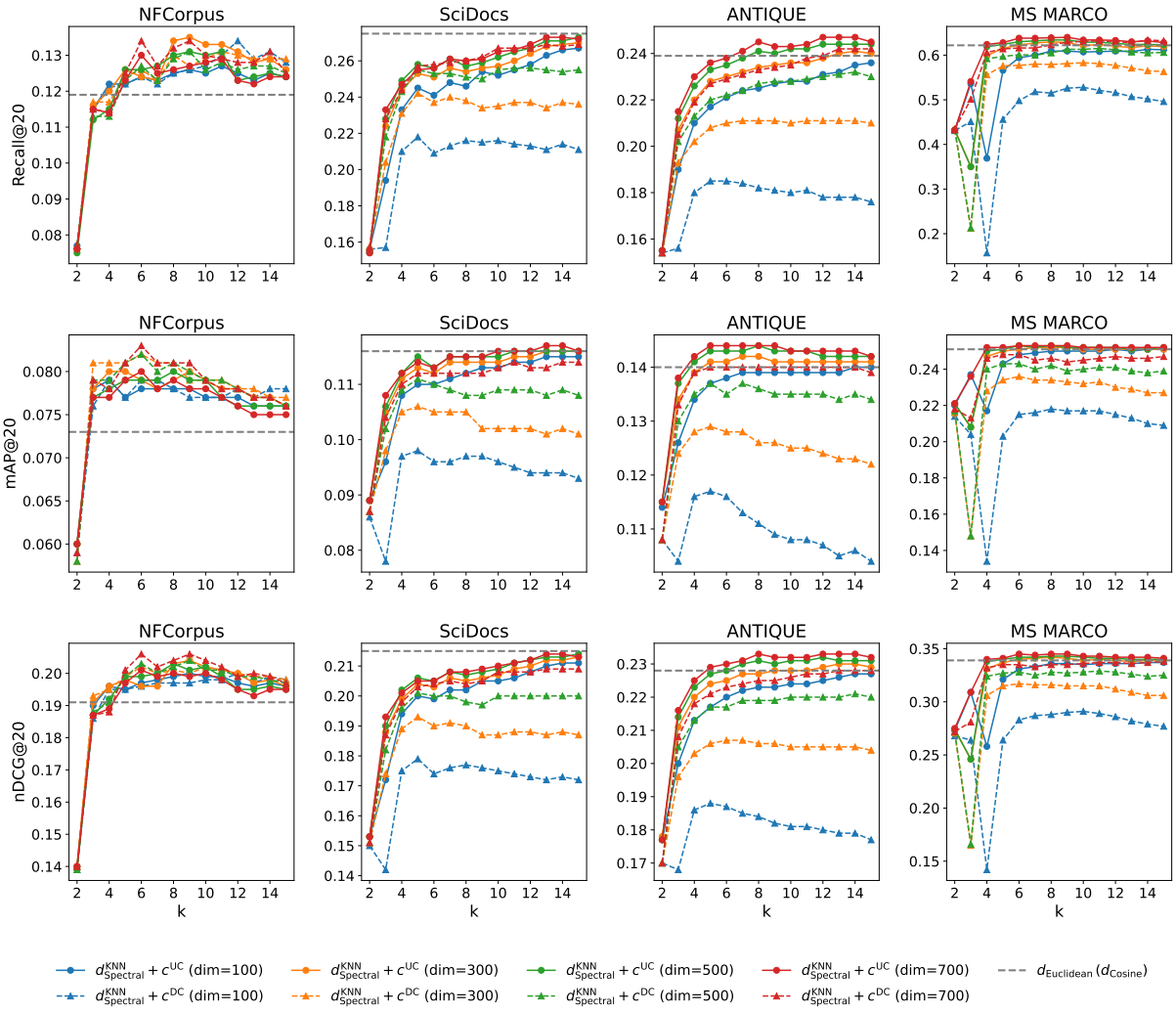

Figure 8: Recall@20, mAP@20 and nDCG@20 comparison of $d^{\text{KNN}}_{\text{Spectral}}$ with SciNCL across varying k.