CRAFT: Corpus Relatedness Analysis Using Fourier Transforms

Kaiwen Chen
University Of Toronto
Toronto, Ontario, Canada
kckevinchen@cs.toronto.edu

Nick Koudas
University Of Toronto
Toronto, Ontario, Canada
koudas@cs.toronto.edu

ABSTRACT

A fundamental challenge in data management is the efficient discovery of term relationships from massive, unstructured text corpora—a critical first step in knowledge-graph construction. This discovery task faces prohibitive computational barriers: the quadratic $O(N^2)$ complexity of an all-pairs analysis and the intractability of processing the full term-document matrix. While dimensionality reduction via embeddings offers a partial solution, the resulting vector proximity often captures broad thematic similarity, failing to isolate the precise co-occurrence signals required for high-quality relation extraction.

This paper introduces CRAFT, a system that overcomes these limitations by recasting term-relatedness discovery as a scalable signal-processing problem. CRAFT's methodology decouples the discovery process from both the term-document matrix and quadratic-time comparisons. First, it employs a randomized Fourier transform to sketch term-occurrence signals directly into a low-dimensional complex space, a process that provably preserves the inner products essential for correlation analysis without materializing the underlying matrix. Second, to break the quadratic barrier, CRAFT leverages the inherent sparsity of term relationships by formulating discovery as a compressed-sensing task. This enables the recovery of significant correlations for any given term directly from its compressed sketch via an efficient Orthogonal Matching Pursuit algorithm, obviating the need for an all-pairs comparison. Our end-to-end implementation and comprehensive experimental evaluation show that CRAFT outperforms modern baselines in both efficiency and precision, enabling high-quality relation discovery at a previously infeasible scale.

PVLDB Reference Format:
Kaiwen Chen and Nick Koudas. CRAFT: Corpus Relatedness Analysis Using Fourier Transforms. PVLDB, 19(10): 2658 - 2671, 2026.
doi:10.14778/3828612.3828622

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/kckevinchen/CRAFT.

1 INTRODUCTION

Identifying frequently co-occurring terms in massive, unstructured text corpora is fundamental to modern data management and AI. When terms consistently appear together—a phenomenon we refer to as *relatedness*—they exhibit strong semantic or functional associations arising from shared topics (e.g., "iPhone" and "Apple"), common contexts (e.g., "aspirin" and "blood clot" in medicine), or compound concepts (e.g., "machine learning"). This capability is critical for applications ranging from query expansion and trend analysis to knowledge discovery in scientific literature [10, 17, 46].

A primary application is the construction of Knowledge Graphs (KGs) in novel domains [2, 5, 23]. KGs represent information as networks of entities and relationships, powering semantic search and question-answering systems. They are increasingly vital for grounding Large Language Models (LLMs) [35, 39, 40], serving as verifiable memory sources in Retrieval-Augmented Generation (RAG) [30] frameworks that reduce hallucinations and inject domain-specific facts. However, their utility is constrained by a fundamental computational bottleneck: bootstrapping the graph from raw, unstructured text.

This bottleneck centers on the *candidate discovery* problem of scalably identifying potentially related terms prior to fine-grained analysis. Current academic benchmarks [8, 38] largely sidestep this challenge, focusing instead on relation classification—assigning semantic labels to pre-identified entity pairs [29, 37]. This formulation assumes that a domain-adapted Entity Recognition tool has already identified all relevant entities, effectively bypassing the core difficulty: one cannot leverage a domain-specific KG without first constructing it from the source text.

In real-world scenarios involving novel, domain-specific corpora (scientific literature, financial reports, enterprise documents), this assumption fails. The task becomes a massive-scale filtering problem: evaluating the combinatorially vast space of potential term pairs—which scales quadratically as $O(N^2)$ with vocabulary size N—to isolate the small fraction of pairs that are meaningfully related.

Existing methods make critical trade-offs. Brute-force pairwise analysis (e.g., Pointwise Mutual Information [11], co-occurrence graphs) generates excessive noise and impractically dense outputs. Embedding-based methods [36, 41, 51] are designed to capture semantic similarity—mapping terms with similar meanings (e.g., *physician* and *doctor*) to nearby points in a dense vector space—but proximity in this space does not imply rigorous statistical correlation.

The underlying limitation stems from the distinction between *paradigmatic* and *syntagmatic* relationships. Embedding methods primarily capture paradigmatic similarity, in which terms can substitute for one another in a sentence (e.g., *physician* and *doctor*). Knowledge discovery, however, often requires identifying syntagmatic associations, in which terms physically co-occur to form a functional unit (e.g., *physician* and *prescription*).

Proceedings of the VLDB Endowment, Vol. 19, No. 10 ISSN 2150-8097.
doi:10.14778/3828612.3828622

For example, "aspirin" and "platelet aggregation" are not paradigmatic substitutes, so dense semantic embeddings may place them far apart. Yet they co-occur consistently in clinical documents because aspirin inhibits platelet aggregation; a syntagmatic, correlation-based analysis identifies this relationship directly. More broadly, persistent co-occurrences provide a high-precision candidate set for downstream Relation Extraction or Entity Recognition tools, avoiding an intractable $O(N^2)$ input space.

LLMs excel at focused reasoning but cannot perform exhaustive corpus-wide discovery and are prone to hallucinating unsupported relationships [24]. This gap calls for a scalable "first-pass" filter for candidate discovery—one that complements LLMs in a hybrid pipeline of lightweight discovery at scale followed by economical LLM-based verification using corpus evidence. We introduce **CRAFT** (**C**orpus **R**elatedness **A**nalysis using **F**ourier **T**ransforms), a novel paradigm that addresses this need, as illustrated in Figure 1. We distinguish between inferring latent similarity and measuring observable correlation; CRAFT is designed strictly for the latter. Rather than learning opaque dense vectors, CRAFT treats term presence as a discrete signal and processes it in a compact frequency-domain representation, efficiently identifying persistent co-occurrences and formally relating them to a statistical correlation measure (the Pearson coefficient). This yields a clean, high-precision candidate set well suited to KG bootstrapping, search enrichment, and LLM validation. Our key insight is the *sparsity-of-effects* principle: most terms interact significantly with only a small number of others.

The primary contributions of our work are as follows:[1]

- **A novel signal-processing paradigm for text analysis.** We introduce CRAFT, a framework that reinterprets term-relatedness discovery in large text corpora as a signal-processing task. By treating the presence of terms across documents as discrete signals and analyzing their frequency-domain representations, CRAFT enables efficient and scalable detection of semantic relationships without relying on pre-existing knowledge bases or expensive pairwise computations.
- **Randomized Fourier embedding with theoretical guarantees.** We propose a Random Fourier Transform (RFT)-based dimensionality-reduction technique that projects high-dimensional term vectors into a low-dimensional complex space. This method preserves inner products with high probability, reducing storage and computation from $O(N^2M)$ to $O(Nk \log M)$ with $k \leq N$ (for N terms and M documents).
- **Cross-Power Spectral Density (CPSD) for correlation estimation.** We adapt CPSD—a classical signal-processing tool—to estimate term correlations in the frequency domain. CPSD captures both the magnitude of correlation and, through phase information, distinguishes positive, negative, and orthogonal relationships, providing a richer and more interpretable measure of association.
- **Sparse recovery via compressed sensing.** We reformulate correlation recovery as a compressed-sensing problem, leveraging the insight that most term relationships are sparse. We employ **Orthogonal Matching Pursuit (OMP)** [18, 48], a computationally efficient algorithm, to recover sparse correlation vectors directly from their compressed measurements, avoiding the prohibitive cost of computing all $O(N^2)$ correlations.
- **End-to-end scalable system with statistical rigor.** The full CRAFT pipeline integrates preprocessing, spectral sketching, CPSD-based correlation analysis, and sparse recovery, equipped with non-asymptotic error bounds, FDR-controlled significance testing, and complexity guarantees.
- **Empirical and theoretical superiority.** Through complexity analysis and empirical validation on six datasets and a biomedical case study against KEGG, we show that CRAFT outperforms existing methods in speed and precision at a scale previously infeasible.

[1]Full report available at: https://github.com/kckevinchen/CRAFT

2 PROBLEM FORMULATION

Our goal is to automatically discover meaningful associations between terms by inductively analyzing a massive collection of unstructured text, such as a decade of biomedical preprints from arXiv's q-bio section or a proprietary repository of clinical-trial documents from a pharmaceutical company. Unlike methods that rely on pre-existing ontologies, this approach processes the entire corpus—examining relatedness (co-occurrence) patterns across term pairs—to uncover significant statistical relationships without prior guidance on what to search for.

The aim is to generate a ranked list of strongly associated term pairs, such as (CRISPR, Cas9, 0.98) or (ibuprofen, cyclooxygenase, 0.92), where the numerical score is the Pearson correlation coefficient that reflects the strength of the relationship. This method is especially powerful for revealing non-obvious yet plausible connections—for instance, linking metabolic pathways such as *succinate* signaling to immune receptors such as *GPR91*, or associating *autophagy*-related genes such as *ULK1* with specific cellular processes—effectively building the foundations of a novel knowledge graph directly from raw text.

More formally, our problem is defined as follows:

- **Input:** A corpus of documents $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ and an associated vocabulary of terms $\mathcal{V} = \{w_1, w_2, \ldots, w_N\}$, where N is the vocabulary size. $\mathcal{V}$ is typically derived by tokenizing $\mathcal{D}$.
- **Output:** A ranked list of term pairs $\ell = [(w_i, w_j, s_{ij}), \ldots]$, where (w_i, w_j) is a pair of distinct terms from $\mathcal{V}$ and $s_{ij} \in \mathbb{R}$ is a score representing the strength of their relationship expressed in terms of the Pearson correlation coefficient. The list is ordered by s_{ij} in descending order.

The objective is to generate this list ℓ in a computationally efficient and scalable manner. The methodology for evaluating the overall quality of the generated list ℓ, will be detailed in the experiments section.

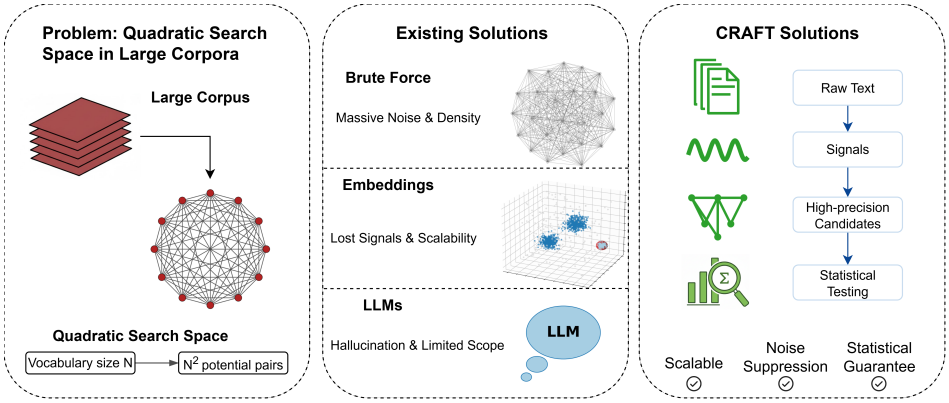


Figure 1: The CRAFT paradigm: term discovery as frequency-domain signal processing, enabling meaningful relationship discovery without exhaustive $O(N^2)$ pairwise comparison.

3 LARGE-SCALE TERM CORRELATION DETECTION VIA RANDOMIZED NONLINEAR EMBEDDING AND SPECTRAL ANALYSIS

This section details our methodology for efficiently detecting statistically significant correlations between all pairs of terms in a large corpus. We first formalize the problem using a term-document matrix, then describe a randomized dimensionality-reduction technique with theoretical guarantees, and finally present our algorithms for correlation estimation, sparse recovery, and significance testing. Detailed proofs are deferred to our technical report [1].

3.1 Term-Document Matrix Construction and Statistical Formulation

We represent the corpus as a term-document matrix $\mathbf{A} \in \mathbb{R}^{N \times M}$, where each element a_{ij} quantifies the importance of term w_i in document d_j. We employ the BM25 weighting function [44],[2] a widely used ranking function that is robust to document-length variation:

$$a_{ij} = \frac{\mathrm{tf}(w_i, d_j) \cdot (k_1 + 1)}{\mathrm{tf}(w_i, d_j) + k_1 \cdot \left(1 - b + b \cdot \frac{|d_j|}{\mathrm{avgdl}}\right)} \cdot \log\left(\frac{M + 1}{\mathrm{df}(w_i) + 0.5}\right) \quad (1)$$

where $\mathrm{tf}(w_i, d_j)$ is the term frequency of w_i in d_j, $|d_j|$ is the length of document d_j, avgdl is the average document length in $\mathcal{D}$, $\mathrm{df}(w_i)$ is the document frequency of w_i, and $k_1 = 1.2$, $b = 0.75$ are standard tuning parameters.

To analyze correlation, we center each row of $\mathbf{A}$ to obtain a matrix of deviation vectors, $\tilde{\mathbf{A}}$. The centered value $\tilde{a}_{ij}$ for term w_i in document d_j is given by $\tilde{a}_{ij} = a_{ij} - \mu_i$, where $\mu_i = \frac{1}{M}\sum_{m=1}^{M} a_{im}$ is the mean score for term w_i. The Pearson correlation coefficient ρ_{ij} between terms w_i and w_j is then defined as the cosine similarity between their centered vectors $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{a}}_j$:

$$\rho_{ij} = \frac{\langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle}{\|\tilde{\mathbf{a}}_i\|_2 \|\tilde{\mathbf{a}}_j\|_2} = \frac{\sum_{m=1}^{M} \tilde{a}_{im}\tilde{a}_{jm}}{\sqrt{\sum_{m=1}^{M} \tilde{a}_{im}^2}\sqrt{\sum_{m=1}^{M} \tilde{a}_{jm}^2}} \quad (2)$$

The direct computation of the full $N \times N$ correlation matrix requires $O(N^2M)$ operations and $O(NM)$ storage, which is computationally infeasible for large N and M.

3.2 Randomized Fourier Embedding with Energy Preservation Guarantees

To overcome this limitation, we employ a randomized nonlinear embedding inspired by Random Fourier Features (RFF) [42]. We project the high-dimensional, centered term vectors $\tilde{\mathbf{a}}_i$ into a substantially lower-dimensional complex space $\mathbb{C}^k$ while *provably preserving* the inner-product information between vectors.

Formally, for each centered term vector $\tilde{\mathbf{a}}_i \in \mathbb{R}^M$, we compute its low-dimensional complex representation $\mathbf{b}_i \in \mathbb{C}^k$ via the following linear transformation:

$$\mathbf{b}_i = \frac{1}{\sqrt{k}} \mathbf{\Phi} \tilde{\mathbf{a}}_i \quad (3)$$

The crucial element of this embedding is the *random Fourier matrix* $\mathbf{\Phi} \in \mathbb{C}^{k \times M}$. The entries of $\mathbf{\Phi}$ are not learned from data but are generated stochastically. Each element is defined by:

$$\phi_{lm} = e^{-2\pi i \xi_l m / M} \quad \text{for} \quad l = 1, \ldots, k \quad \text{and} \quad m = 1, \ldots, M, \quad (4)$$

where each frequency parameter ξ_l is sampled independently and identically from a Uniform$(0, M)$ distribution.

Theoretical Underpinnings and Guarantees. The quality of this embedding is established through concentration-of-measure arguments in the spirit of the Johnson-Lindenstrauss (JL) lemma [27], without requiring kernel-theoretic machinery such as Bochner's theorem (which applies only to shift-invariant kernels and is therefore inapplicable to the linear kernel used here). Each component of the inner product $\mathbf{b}_i^* \mathbf{b}_j$ is an unbiased estimate of $\tilde{\mathbf{a}}_i^\top \tilde{\mathbf{a}}_j$, and because the k frequencies are drawn independently, standard concentration inequalities give convergence to the true inner product at rate $O(1/\sqrt{k})$, matching the ϵ-distortion guarantees of classical JL projections. The full proof is provided in our technical report [1].

[2] Any standard term-weighting scheme (e.g., TF-IDF, log-frequency, or PPMI) is compatible with the framework.

The following theorem establishes the key properties of the estimator formed by these embeddings.

Theorem 3.1 (Unbiased Inner Product Estimator). *Let $\mathbf{b}_i$ and $\mathbf{b}_j$ be the randomized Fourier embeddings of $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{a}}_j$, respectively, as defined by Equation 3. Then, the complex inner product $\langle \mathbf{b}_i, \mathbf{b}_j \rangle = \mathbf{b}_i^* \mathbf{b}_j$ is an unbiased estimator of the true inner product:*

$$\mathbb{E}_{\Phi}[\mathbf{b}_i^* \mathbf{b}_j] = \tilde{\mathbf{a}}_i^\top \tilde{\mathbf{a}}_j. \tag{5}$$

Furthermore, the variance of the estimator decays linearly with the embedding dimension k:

$$\mathrm{Var}(\mathbf{b}_i^* \mathbf{b}_j) = O(1/k). \tag{6}$$

Unbiasedness ensures that downstream analyses converge in expectation, while the $O(1/\sqrt{k})$ concentration rate keeps approximation error tightly bounded for a modest $k \sim O(\log M)$, reducing the cost of materializing pairwise inner products from $O(N^2 M)$ to $O(Nk)$ ($k \ll M, N$).

A consequence of this random projection is that the Johnson-Lindenstrauss lemma holds. For a target dimension $k = O(\epsilon^{-2} \log M)$, all pairwise inner products are preserved within an ϵ-factor with high probability:

$$\mathbb{P}\left(\left|\langle \mathbf{b}_i, \mathbf{b}_j \rangle - \langle \tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j \rangle\right| \geq \epsilon \|\tilde{\mathbf{a}}_i\| \|\tilde{\mathbf{a}}_j\|\right) \leq 2e^{-k\epsilon^2/4} \tag{7}$$

Computation and storage. The operation $\tilde{\mathbf{a}}_i \mathbf{\Phi}^\top$ is a random Fourier transform of $\tilde{\mathbf{a}}_i$, computable in $O(k \log M)$ time per term via the Non-Uniform FFT (NUFFT) [20], for a total of $O(Nk \log M)$. The resulting embedding matrix $\mathbf{B} \in \mathbb{C}^{N \times k}$ requires only $O(Nk)$ storage—a dramatic reduction from $O(NM)$ since $k \ll M$ (e.g., $k \sim 10^3$–10^4 while M can be in the billions).

3.3 Cross-Power Spectral Density Estimation with Phase Analysis

Having projected the high-dimensional, centered term vectors $\tilde{\mathbf{a}}_i$ into the low-dimensional complex space via the randomized Fourier embedding $\mathbf{b}_i = \frac{1}{\sqrt{k}} \mathbf{\Phi} \tilde{\mathbf{a}}_i$, we now construct a similarity matrix in this embedded space. We define the *Cross-Power Spectral Density (CPSD)* matrix $\mathbf{P} \in \mathbb{C}^{N \times N}$ for the embedded terms [3]. The entries of $\mathbf{P}$ are given by the complex inner products of the embeddings:

$$P_{ij} = \langle \mathbf{b}_i, \mathbf{b}_j \rangle = \mathbf{b}_i^* \mathbf{b}_j \tag{8}$$

This matrix is Hermitian ($\mathbf{P} = \mathbf{P}^*$) and serves as a statistically well-founded proxy for the matrix $\mathbf{G} = \tilde{\mathbf{A}} \tilde{\mathbf{A}}^\top$.

Theoretical Justification: From CPSD to Correlation. The following theorem formalizes the connection between the CPSD matrix and the desired correlation coefficients, providing the core justification for our approach.

Theorem 3.2 (CPSD Correlation Estimator). *Let $\mathbf{b}_i$ and $\mathbf{b}_j$ be the randomized Fourier embeddings of the centered term vectors $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{a}}_j$, respectively. Let $P_{ij} = \mathbf{b}_i^* \mathbf{b}_j$ be their Cross-Power Spectral Density. Then, the estimator:*

$$\hat{\rho}_{ij} = \frac{\Re(P_{ij})}{\sqrt{P_{ii}} \sqrt{P_{jj}}} \tag{9}$$

is a consistent estimator for the true Pearson correlation coefficient ρ_{ij} between terms i and j. That is, $\hat{\rho}_{ij} \xrightarrow{p} \rho_{ij}$ as the embedding dimension $k \to \infty$. $\Re(P_{ij})$ takes the real part of the complex-valued inner product.

The Informative Role of Phase. A significant advantage of operating in the complex plane is the information encoded in the *phase* $\theta_{ij} = \arg(P_{ij})$. Phase near 0 signals positive correlation (embeddings in phase), phase near π signals negative correlation (anti-phase), and phase near $\pm\pi/2$ signals orthogonality ($\Re(P_{ij}) \approx 0$, uncorrelated terms). This geometric interpretation allows qualitative assessment of relationships directly from the complex-valued $\mathbf{P}$ matrix without computing the normalized ratio $\hat{\rho}_{ij}$ for every pair.

Non-Asymptotic Error Bounds. The quality of the estimation is not merely asymptotic; it comes with strong, non-asymptotic probabilistic guarantees crucial for applications.

Lemma 3.3 (Error Bound for Correlation Estimation). *For any pair (i, j), any embedding dimension k, and any confidence parameter $\delta \in (0, 1)$, the error of the correlation estimator is bounded with high probability:*

$$\mathbb{P}\left(|\hat{\rho}_{ij} - \rho_{ij}| \leq C\sqrt{\frac{\log(1/\delta)}{k}}\right) \geq 1 - \delta \tag{10}$$

where C is a constant independent of N and M.

This bound, derivable from standard concentration inequalities, confirms an $O(1/\sqrt{k})$ error decay and yields a principled rule for selecting k: solving $C\sqrt{\log(1/\delta)/k} \leq \epsilon$ gives $k \geq C^2 \log(1/\delta)/\epsilon^2$. The logarithmic dependence on δ and inverse-quadratic dependence on ϵ are what make this randomized approach scalable in practice.

Algorithmic Implication and Complexity. Although the CPSD matrix $\mathbf{P} = \mathbf{B}\mathbf{B}^*$ provides the theoretical foundation for correlation estimation, its explicit construction is itself an $O(N^2 k)$ bottleneck, infeasible when N reaches the millions. The next section shows how to bypass this cost via sparse recovery, accurately estimating the correlations of each term without ever materializing $\mathbf{P}$.

4 SPARSE CORRELATION RECOVERY VIA COMPRESSED SENSING

A central challenge in our setting is efficiently discovering correlations between a term and all other terms in the dataset. The naive approach of computing all pairwise correlations is computationally prohibitive for large N, scaling as $O(N^2)$. A key insight underpinning our method is the *sparsity-of-effects* principle: in most cases, a given term w_i interacts significantly with only a small number $S \ll N$ of other terms. The correlation vector $\boldsymbol{\rho}_i = (\rho_{i1}, \ldots, \rho_{iN})^\top$ is therefore S-**sparse**, or at least approximately so.

We exploit this sparsity by reformulating correlation recovery as a **compressed-sensing (CS)** task [9, 45]. Instead of measuring all N potential correlations directly, we acquire a small number $k \ll N$ of **compressed measurements** and recover the sparse correlation vector through a sparse-recovery algorithm.

[3] This matrix will not be computed nor materialized in the sequel.

4.1 Sparse Recovery via Orthogonal Matching Pursuit

For each term w_i, we recover its correlation vector using Orthogonal Matching Pursuit (OMP) [18, 48], an iterative greedy algorithm that offers a computationally efficient alternative to convex optimization. Rather than solving a single minimization problem, OMP greedily constructs the sparse solution vector $\mathbf{x}_i \in \mathbb{R}^N$ over S steps, where S is the target sparsity.

The algorithm maintains a residual vector $\mathbf{r}$, initialized as the measurement vector $\mathbf{z}_i$. In each step, it performs two key operations:

(1) **Identification:** It searches for the column $\boldsymbol{\psi}_j$ in the sensing matrix $\boldsymbol{\Psi}$ that is most correlated with the current residual $\mathbf{r}$.
(2) **Projection:** It adds this column to an active set and then calculates the least-squares solution for the coefficients of all currently active columns, ensuring an optimal fit at each step. The residual is then updated by subtracting this new fit.

4.2 Sensing Matrix and Measurement Vector

The efficacy of the compressed sensing approach critically depends on the design of the **sensing matrix** $\boldsymbol{\Psi}$ and the **measurement vector** $\mathbf{z}_i$.

The sensing matrix $\boldsymbol{\Psi} \in \mathbb{C}^{k \times N}$ is constructed from normalized measurement vectors (computed in Section 3.2) and is formally defined as:

$$\boldsymbol{\Psi} = \left[\frac{\mathbf{b}_1}{\sqrt{P_{11}}}, \frac{\mathbf{b}_2}{\sqrt{P_{22}}}, \ldots, \frac{\mathbf{b}_N}{\sqrt{P_{NN}}} \right]^\top. \tag{11}$$

Each column j of $\boldsymbol{\Psi}$ corresponds to a term w_j and is given by $\mathbf{b}_j/\sqrt{P_{jj}}$. The normalization by $\sqrt{P_{jj}}$, which approximates the standard deviation of term w_j's vector,[4] is crucial as it ensures that the energy of each column is controlled—a necessary condition for the theoretical recovery guarantees. This matrix acts as a linear projection operator that maps the high-dimensional correlation vector $\boldsymbol{\rho}_i \in \mathbb{R}^N$ down to the low-dimensional measurement space $\mathbb{C}^k$.

The corresponding measurement vector for term w_i is defined as $\mathbf{z}_i = \mathbf{b}_i/\sqrt{P_{ii}} \in \mathbb{C}^k$. This vector represents the observed compressed measurement. The entire system is designed such that this observation approximates a linear combination of the correlations ρ_{ij} between w_i and all other terms w_j, sensed through the matrix $\boldsymbol{\Psi}$:

$$\mathbf{z}_i \approx \boldsymbol{\Psi}\boldsymbol{\rho}_i = \sum_{j=1}^{N} \rho_{ij} \cdot \frac{\mathbf{b}_j}{\sqrt{P_{jj}}}. \tag{12}$$

This approximation contains noise and errors inherent in the embedding process, which we can model as an error vector $\mathbf{e}$ such that $\mathbf{z}_i = \boldsymbol{\Psi}\boldsymbol{\rho}_i + \mathbf{e}$. A standard result in compressed sensing theory is that the expected energy of this stochastic error, $\|\mathbf{e}\|_2$, scales with the number of measurements as $O(\sqrt{k})$ [18]. The OMP algorithm is robust to this error, which is explicitly accounted for in its theoretical guarantees.

[4] b_j is zero mean and $\sigma_j^2 = \frac{1}{k}\sum_{i=1}^{k} |b_{ji}|^2 = P_{jj}$

4.3 Theoretical Guarantee: Stable Sparse Recovery

The power of this greedy approach is justified by the following theorem, which provides a rigorous bound on the recovery error for OMP.

Theorem 4.1 (OMP Recovery Guarantee). *If the sensing matrix $\boldsymbol{\Psi}$ satisfies the **Restricted Isometry Property (RIP)** of order $S+1$ with a sufficiently small constant δ_{S+1}, then the solution $\hat{\mathbf{x}}_i$ produced by the OMP algorithm after S steps satisfies:*

$$\|\hat{\mathbf{x}}_i - \boldsymbol{\rho}_i\|_2 \leq C_1\|\mathbf{e}\|_2 + C_2\frac{\|\boldsymbol{\rho}_i - \boldsymbol{\rho}_i^S\|_1}{\sqrt{S}} \tag{13}$$

where $\mathbf{e}$ is the noise vector from the measurement model, $\boldsymbol{\rho}_i^S$ is the best S-term approximation of $\boldsymbol{\rho}_i$, and C_1, C_2 are constants.

Interpretation. The RIP condition requires that $\boldsymbol{\Psi}$ approximately preserves the lengths of all sparse vectors, ensuring OMP's greedy selections are reliable; random Fourier-based matrices satisfy the RIP with high probability when $k = O(S\log(N/S))$ [6, 9]. The error bound decomposes into a *noise term* $C_1\|\mathbf{e}\|_2$, which scales as $O(\sqrt{k})$, and an *approximation term* $C_2\|\boldsymbol{\rho}_i - \boldsymbol{\rho}_i^S\|_1/\sqrt{S}$, which vanishes for exactly S-sparse vectors and remains small for compressible ones, ensuring graceful degradation.

4.4 Practical Implementation

The core computational challenge is executing OMP for each term. Naively running OMP against all N other terms is prohibitive, carrying a cost of at least $O(N^2Sk)$. To achieve tractability, we first precompute an approximate L-nearest-neighbor (L-NN) graph in the embedding space [19, 31, 52, 55], which restricts the recovery problem for each term to a small candidate set of $L \ll N$ neighbors.

A simpler alternative—computing direct dot products with the L neighbors (the CRAFT-DP baseline)—is computationally efficient but has two critical drawbacks that OMP overcomes:

- **Denoising:** OMP's model-based recovery ($\mathbf{z}_i \approx \boldsymbol{\Psi}\boldsymbol{\rho}_i$) filters embedding noise, retaining only the most signal-rich candidates.
- **Non-redundancy:** OMP considers all candidates simultaneously, "explaining away" redundant correlates and revealing a parsimonious set of drivers.

For OMP applied to the restricted set of L neighbors, the total cost reduces to a tractable $O(NSLk)$, which is linear in the number of terms N. The final output is a sparse graph requiring $O(NL)$ storage. To empirically validate these claims, we present a comparison of CRAFT against CRAFT-DP in Section 6.

4.5 Statistical Significance Testing and Multiple Testing Correction

In large-scale correlation analyses, the sheer number of pairwise tests performed—on the order of $O(N^2)$—inevitably leads to a high number of false positives if statistical significance is not rigorously assessed. Without proper correction, many spurious correlations may be mistakenly identified as meaningful, undermining the reliability of the results. Therefore, we incorporate a two-stage statistical testing procedure: first, we assess the significance of each individual

correlation estimate, and second, we apply a multiple testing correction to control the overall false discovery rate across the entire set of hypotheses.

For each estimated correlation $\hat{\rho}_{ij}$, we test the null hypothesis $H_0 : \rho_{ij} = 0$. We apply Fisher's z-transform to transform the raw estimates into a statistic and stabilize the variance of the correlation coefficient:

$$z_{ij} = \frac{1}{2} \log \left(\frac{1 + \hat{\rho}_{ij}}{1 - \hat{\rho}_{ij}} \right) \tag{14}$$

Under H_0, the statistic z_{ij} is approximately normally distributed with mean 0 and variance $\frac{1}{k-3}$ (where k is the embedding dimension). We use k rather than the corpus size M as the degrees of freedom because $\hat{\rho}_{ij}$ is estimated from the inner product of two k-dimensional embeddings, each dimension corresponding to an independently sampled random frequency. The embedding dimension k therefore governs the effective number of independent measurements underlying the estimate; using M would overstate the available information and produce artificially narrow confidence intervals. In practice, $k = 256$ yields a more conservative significance test than $M \geq 4{,}000$, reducing false discoveries at the cost of slightly lower statistical power—a desirable trade-off for discovery-oriented applications. The resulting test statistic $T_{ij} = z_{ij}\sqrt{k-3}$ therefore follows a standard normal distribution, $T_{ij} \sim \mathcal{N}(0, 1)$. This allows us to compute two-tailed p-values in the standard way.

Given the $O(N^2)$ hypotheses being tested, a multiple testing correction is imperative to control the false positive rate. We control the False Discovery Rate (FDR) using the Benjamini-Hochberg procedure [7]. For a fixed term w_i, let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ be the sorted p-values of all its m tested correlations. We find the largest index r such that:

$$p_{(r)} \leq \frac{r}{m}\alpha \tag{15}$$

We then reject all hypotheses corresponding to $p_{(1)}, \ldots, p_{(r)}$, where α is the desired significance level (e.g., 0.05). The threshold $p_{\text{BH}} = p_{(r)}$ serves as the corrected significance level for term w_i.

The final output of our algorithm is a set of significant term pairs that meet both a strength and a significance criterion:

$$E = \left\{ (w_i, w_j) : |\hat{\rho}_{ij}| > \tau \quad \text{and} \quad p_{ij} \leq p_{\text{BH}} \right\} \tag{16}$$

where τ is a correlation strength threshold.

Complexity Analysis. The complexity of the algorithm is dominated by the initial signal embedding stage. The randomized Fourier transform takes $O(Nk \log M)$. The subsequent stages are computationally less expensive. The sparse correlation recovery via OMP is performed for each of the N terms on a restricted candidate set of size L. With a target sparsity of S, this recovery has a total time complexity of $O(NSLk)$.

The final statistical validation phase consists of two steps. First, calculating a p-value using the Fisher z-transform for each of the (at most) S non-zero correlations per term takes $O(NS)$ time in total. Second, applying the Benjamini-Hochberg procedure requires sorting these S p-values for each of the N terms, resulting in a total cost of $O(NS \log S)$.

Algorithm 1 The End-to-End CRAFT Pipeline

1: **Input:** Corpus $\mathcal{D}$, Embedding dim. k, Neighbor set size L, OMP sparsity S, Significance level α, Correlation threshold τ
2: **Output:** Ranked list of significant scored pairs ℓ
3: $\mathcal{V} \leftarrow \text{BuildVocabulary}(\mathcal{D})$
4: $\mathbf{B} \leftarrow \text{zeros}(N, k, \text{dtype=complex})$
5: **for** each term $w_i \in \mathcal{V}$ (from $i = 1$ to N) **do**
6: $\quad \tilde{\mathbf{a}}_i \leftarrow \text{GenerateCenteredSignal}(\mathcal{D}, w_i)$
7: $\quad \mathbf{b}_i \leftarrow \text{ApplyRFT}(\tilde{\mathbf{a}}_i)$
8: $\quad \mathbf{B}[i, :] \leftarrow \mathbf{b}_i^\top$
9: **end for**
10: $\mathbf{B}_{\text{norm}} \leftarrow \text{NormalizeRows}(\mathbf{B})$
11: $\text{FaissIndex} \leftarrow \text{BuildHNSWIndex}(\mathbf{B}_{\text{norm}})$
12: $\mathcal{L} \leftarrow \text{new map}()$
13: **for** each term $w_i \in \mathcal{V}$ (from $i = 1$ to N) **do**
14: $\quad \mathcal{L}_i \leftarrow \text{FaissIndex.Search}(\mathbf{B}_{\text{norm}}[i, :], L)$
15: $\quad \mathcal{L}[i] \leftarrow \mathcal{L}_i$
16: **end for**
17: $E \leftarrow \emptyset$
18: $\mathbf{\Psi} \leftarrow \mathbf{B}_{\text{norm}}^\top$
19: **for** each term $w_i \in \mathcal{V}$ (from $i = 1$ to N) **do**
20: $\quad \mathcal{L}_i \leftarrow \mathcal{L}[i]$
21: $\quad \mathbf{z}_i \leftarrow \mathbf{\Psi}[:, i]$
22: $\quad \mathbf{\Psi}_{\mathcal{L}_i} \leftarrow \mathbf{\Psi}[:, \mathcal{L}_i]$
23: $\quad \hat{\mathbf{x}}_i^{\mathcal{L}} \leftarrow \text{SolveOMP}(\mathbf{\Psi}_{\mathcal{L}_i}, \mathbf{z}_i, S)$
24: $\quad \text{p\_values} \leftarrow []$
25: $\quad \text{results} \leftarrow []$
26: $\quad$ **for** each non-zero $\hat{\rho}_{ij}$ in $\hat{\mathbf{x}}_i^{\mathcal{L}}$ (with index $j \in \mathcal{L}_i$) **do**
27: $\quad\quad p_{ij} \leftarrow \text{FisherZTest}(\hat{\rho}_{ij}, k)$
28: $\quad\quad \text{p\_values.append}(p_{ij})$
29: $\quad\quad \text{results.append}((i, j, \hat{\rho}_{ij}, p_{ij}))$
30: $\quad$ **end for**
31: $\quad p_{\text{BH}} \leftarrow \text{BenjaminiHochberg}(\text{p\_values}, \alpha)$
32: $\quad$ **for** $(i, j, \hat{\rho}_{ij}, p_{ij}) \in \text{results}$ **do**
33: $\quad\quad$ **if** $|\hat{\rho}_{ij}| > \tau$ and $p_{ij} \leq p_{\text{BH}}$ **then**
34: $\quad\quad\quad E \leftarrow E \cup \{(w_i, w_j, \hat{\rho}_{ij})\}$
35: $\quad\quad$ **end if**
36: $\quad$ **end for**
37: **end for**
38: $\ell \leftarrow \text{SortByStrength}(E)$
39: **return** ℓ

5 THE CRAFT SYSTEM FRAMEWORK

We now present the end-to-end **CRAFT** system. As illustrated in Figure 2 and formalized in Algorithm 1, CRAFT is organized as a scalable three-stage pipeline that transforms a raw text corpus into a validated set of statistically significant term pairs.

Stage 1: Pipelined Signal Generation and Embedding. The first stage (lines 3–9 of Algorithm 1) transforms the corpus into a dense embedding matrix $\mathbf{B} \in \mathbb{C}^{N \times k}$. For each term w_i, CRAFT generates its BM25-weighted, mean-centered signal $\tilde{\mathbf{a}}_i$ and immediately projects it into a k-dimensional complex embedding $\mathbf{b}_i$ via the Randomized Fourier Transform (Equation (3)), which provably preserves inner products (Theorem 3.1). Because each signal is embedded on the

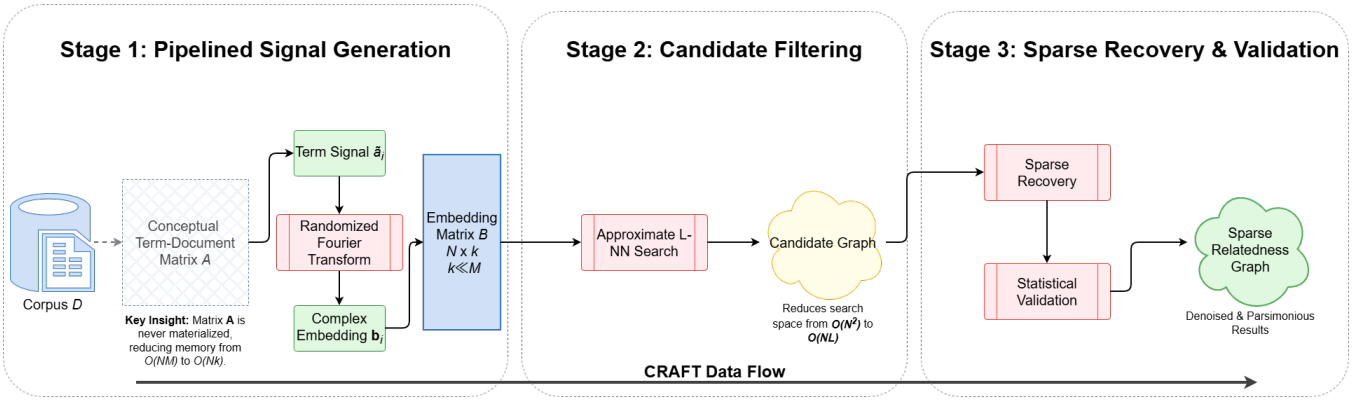


Figure 2: The CRAFT system architecture, illustrating the three main stages: (1) pipelined signal generation and embedding, (2) candidate filtering via approximate k-NN, and (3) sparse recovery and statistical validation.

fly, the conceptual $N \times M$ term-document matrix is **never explicitly materialized**, reducing the memory footprint from $O(NM)$ to $O(Nk)$ and enabling processing of massive corpora on commodity hardware.

Stage 2: Candidate Filtering via Approximate k-NN. A naive $O(N^2k)$ all-pairs comparison is intractable. Exploiting the sparsity-of-effects hypothesis, we instead build a Hierarchical Navigable Small World (HNSW) index [33] on the normalized embeddings using Faiss [15] (lines 10–16). For each of the N terms, we retrieve its $L \ll N$ nearest neighbors, yielding a candidate graph that reduces the search space from $O(N^2)$ to $O(NL)$.

Stage 3: Sparse Recovery and Statistical Validation. The final stage (lines 17–40) recovers sparse, high-quality correlations for each term via Orthogonal Matching Pursuit (OMP). For each term w_i, OMP solves $\mathbf{z}_i \approx \mathbf{\Psi}_{\mathcal{L}_i}\mathbf{x}_i$, where $\mathbf{z}_i$ is the normalized embedding and $\mathbf{\Psi}_{\mathcal{L}_i}$ contains the embeddings of its L candidates. OMP iteratively identifies the candidate most correlated with the current residual and updates the least-squares fit, producing a sparse correlation vector $\hat{\mathbf{x}}_i^{\mathcal{L}}$. The recovered correlations then undergo statistical validation: p-values are computed via Fisher's z-transform, and the Benjamini–Hochberg procedure controls the false discovery rate at level α. Only pairs passing both a strength threshold and the corrected significance level enter the final ranked list ℓ. Since each term is processed independently on its restricted candidate set, this stage is embarrassingly parallel.

6 EXPERIMENTAL EVALUATION

In this section, we empirically evaluate CRAFT along four axes: **effectiveness** (quality of discovered correlations vs. baselines), **efficiency and scalability** (runtime and memory as corpus and vocabulary grow), **ablation** (component contributions and hyperparameter sensitivity), and **external validation** (performance on noisy real-world text against an independently curated ground truth and a downstream application).

6.1 Experimental Setup

Our experimental design rigorously evaluates CRAFT's performance along two primary dimensions: **effectiveness** and **efficiency**. For effectiveness, we evaluate all methods under two scenarios: (1) a **ranked retrieval** task in which the generated candidate pairs are ordered by similarity score, and (2) a **significance set** task in which a statistical test is used to select a final set of pairs. To support this evaluation, we employ two categories of datasets.

6.1.1 Benchmark Datasets for Effectiveness Evaluation. To evaluate effectiveness, we employ two standard knowledge-graph construction benchmarks from the Text2KG suite [43], both of which provide ground-truth triples for each text document. Their well-structured nature and moderate scale are well suited to a precise quantitative and qualitative analysis of the discovered correlations.

- **Wikidata-TekGen:** A dataset of 13K documents synthesized from Wikidata triples. It is characterized by a broad, general-domain vocabulary of entities and relations.
- **DBpedia-WebNLG:** A corpus of 4K documents containing descriptive, well-structured sentences generated from DBpedia triples, providing a clean and varied evaluation setting.

6.1.2 Large-Scale Corpora for Scalability Evaluation. To address efficiency and scalability, we also test our approach on two large corpora that are representative of challenging, large-scale text. Their significant size allows us to thoroughly test the performance of CRAFT.

- **GenWiki:** A large-scale, general-domain dataset constructed from English Wikipedia [26], comprising over 700K articles with golden triplets for KG construction evaluation.
- **arXiv Corpus:** To evaluate performance on domain-specific scientific text at scale, we use a series of corpora constructed from text chunks extracted from full-text computer science papers from arXiv [12]. Characterized by a specialized vocabulary, these serve as a robust benchmark. We use three versions of increasing size: **arXiv-1M** (1 million text chunks), **arXiv-3M** (3 million chunks), and **arXiv-5M** (5 million chunks).

Table 1: Effectiveness on all datasets, measured by Mean Average Precision (mAP) and F1-Score.

	Wikidata		DBpedia		GenWiki		arXiv-1M		arXiv-3M		arXiv-5M	
Method	**mAP**	**F1**	**mAP**	**F1**	**mAP**	**F1**	**mAP**	**F1**	**mAP**	**F1**	**mAP**	**F1**
RP-ANN	92.7	46.1	97.2	51.2	80.5	42.5	75.3	40.1	74.1	39.5	73.5	39.0
SVD-ANN	95.1	33.6	99.1	40.1	81.2	40.8	77.8	38.4	76.5	37.9	75.9	37.1
CRAFT-DP	99.3	80.1	99.2	78.5	83.6	80.6	82.1	78.2	80.9	77.8	80.5	77.1
CRAFT	98.7	80.5	98.1	80.3	85.0	81.3	80.9	80.5	81.1	79.9	80.1	79.2

6.1.3 Compared Methods. To rigorously evaluate our framework, we compare the full CRAFT pipeline against a series of methods. To ensure a fair comparison of the embedding and recovery stages, all methods leverage the same Approximate Nearest Neighbor (ANN) search component [19, 31, 52] for candidate filtering. The following methods are evaluated:

(1) **RP-ANN:** This method represents a standard, scalable heuristic. It first builds a term-document matrix with BM25 weights, applies a **Random Projection (RP)** [3] to reduce dimensionality, and then performs the ANN search.
(2) **SVD-ANN:** This baseline applies a truncated **Singular Value Decomposition (SVD)** to the BM25 matrix (similar to the classic Latent Semantic Analysis method [13]), to generate dense embeddings for the subsequent ANN search.
(3) **CRAFT-DP:** This variant first applies the CRAFT Fourier embedding and ANN filtering stages. It then isolates our final stage by replacing the OMP sparse recovery step with a simple dot product ranking on the candidate set, allowing us to quantify the benefit of OMP as part of our ablation study.
(4) **CRAFT:** This is our full, end-to-end proposed framework, including the initial Fourier embedding, the ANN candidate filtering, and the final sparse correlation recovery via Orthogonal Matching Pursuit.

6.1.4 Evaluation Metrics & Implementation.

- **Effectiveness:** We evaluate the quality of the discovered term pairs under the two scenarios based on the output format:
  - **Ranked List Evaluation:** For the ranked retrieval task, we report **Mean Average Precision (mAP)**. This metric is the mean of the Average Precision (AP) scores over a set of queries Q, defined as:

$$\text{mAP} = \frac{1}{|Q|} \sum_{q \in Q} \text{AP}(q)$$

    where $\text{AP}(q)$ for a single query is the average of precision values at the position of each correct item in the ranked list. A higher mAP score indicates better performance, as it rewards models for placing correct items at the top of the list.
  - **Significance Set Evaluation:** For the significance set task, we report **F1-Score**.

  A rigorous evaluation of correlation-based discovery methods requires a ground truth that is methodologically aligned with the chosen statistical metric. Accordingly, we construct our benchmark to consist exclusively of pairs that exhibit a strong statistical association as measured by the Pearson correlation coefficient.

  To this end, we operationally define a strong correlation as any pair with a Pearson score at or above the 95th percentile among all the possible pairs. This threshold is applied universally to ensure consistency:
  - For datasets providing a golden set, we perform a refinement, retaining only those golden pairs that meet our 95th percentile criterion.
  - For datasets without a golden set, this threshold is used to construct the ground truth directly.
- **Efficiency:** To characterize system performance, we measure both end-to-end wall-clock time and peak memory usage. Our analysis focuses on runtime as the primary efficiency metric, as the peak memory bottleneck is consistent across all methods.[5] This bottleneck is the storage of the final dense embedding matrix of size $N \times k$, where N is the vocabulary size and k is the embedding dimension. Since this data structure is a shared requirement for all evaluated algorithms, we adopt runtime scalability as the primary axis for our efficiency evaluation.
- **Implementation:** All experiments were conducted on a server equipped with an AMD EPYC 7C13 CPU and 512 GB of RAM. Our system is implemented in Python, utilizing NumPy, SciPy, and the Faiss library for its HNSW index. For OMP, we adopted the fast algorithm from [54]. Unless otherwise noted, the default parameters for CRAFT are an embedding dimension of $k = 256$, a neighbor set size of $L = 100$, an OMP sparsity budget of $S = 10$, and a significance level $\alpha = 0.05$.

6.2 Effectiveness Evaluation

We evaluate effectiveness across all six datasets, with the primary results presented in Table 1. The analysis is divided into two parts, corresponding to our two evaluation scenarios: ranking quality and the quality of the final significance set.

6.2.1 Ranking Quality. As shown in Table 1, our CRAFT variants consistently outperform the baselines. On well-structured benchmarks (**Wikidata**, **DBpedia**), all methods achieve high mAP, with CRAFT-DP reaching a near-perfect 99.3 on Wikidata. The

[5] For SVD-ANN, the full term-document matrix can be prohibitively large if densified. Standard implementations, however, typically operate directly on a sparse representation of this matrix to manage memory consumption.

gap widens on noisier, large-scale corpora: on **arXiv-5M**, CRAFT achieves an mAP of 80.1 versus 75.9 for the strongest baseline, confirming robust ranking quality on real-world text.

A consistent pattern in Table 1 is the divergence between **CRAFT** and **CRAFT-DP**. On the mAP metric, CRAFT-DP occasionally edges ahead—for instance, 99.3 vs. 98.7 on Wikidata and 82.1 vs. 80.9 on arXiv-1M—yet on F1, the full CRAFT model outperforms CRAFT-DP across every dataset. This is explained by the role of Orthogonal Matching Pursuit (OMP) in CRAFT's final stage. OMP acts as a parsimonious filter, greedily selecting only the minimal set of candidates that best explain the observed compressed measurement and pruning redundant or marginally correlated terms. This yields a precise, non-redundant output set that directly boosts precision and F1. The same sparsification, however, can drop true-positive candidates that carry weaker individual signals or are partially redundant with an already-selected correlate, slightly lowering the rank-sensitive mAP. CRAFT-DP, by contrast, retains all candidates with their raw dot-product scores, preserving a fuller ranking at the cost of admitting more noise into the significance set. In effect, OMP trades a small amount of ranking completeness for a substantial gain in output purity—a trade-off well-suited to KG bootstrapping, where precision matters most.

6.2.2 Significance Set Quality. The F1-Score in Table 1 reveals a dramatic gap: both CRAFT variants outperform the baselines, whose dense, noisy ANN neighborhoods produce excessive false positives. The strong F1 of **CRAFT-DP** shows that the Randomized Fourier Transform embeddings already separate signal from noise effectively; the full **CRAFT** model then applies OMP as a sparse recovery filter that further prunes false positives, yielding state-of-the-art F1-scores across all datasets.

6.3 Efficiency and Scalability

To evaluate the practicality of CRAFT for large-scale applications, we analyze its computational performance on our two largest corpora, **GenWiki** and **arXiv** series. We separately analyze runtime and memory usage to provide a comprehensive picture of the system's efficiency.

We evaluate the end-to-end runtime for all methods. Table 2 reports both absolute wall-clock times and the relative slowdown normalized against the fastest baseline, **RP-ANN**.

SVD-ANN uses a randomized SVD [22], but its matrix factorization core still becomes a bottleneck as vocabulary and document counts grow, explaining the significant slowdown on the larger arXiv corpora. **CRAFT**'s embedding stage relies on a near-linear randomized Fourier transform and its OMP recovery operates only on small candidate sets, so overall runtime scales nearly as well as the lightweight RP-ANN heuristic—demonstrating that CRAFT's effectiveness gains come with minimal computational overhead.

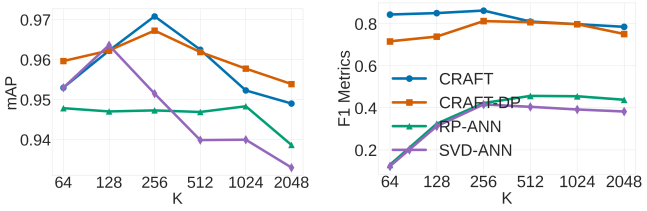
On the largest corpus (arXiv-5M), CRAFT is only 1.23× slower than the lightweight RP-ANN baseline. The scaling trend across the arXiv series is near-linear: from 1M to 5M chunks (a 5× increase in corpus size), CRAFT's runtime grows by a factor of 5.1×, closely tracking RP-ANN (5.0×). In contrast, SVD-ANN exhibits super-linear growth (6.1× over the same range), driven by its reliance on matrix factorization whose cost grows with both vocabulary and document count.

Table 2: Wall-clock runtime (seconds) on the large-scale datasets. Relative slowdown vs. RP-ANN shown in parentheses.

	GenWiki	arXiv-1M	arXiv-3M	arXiv-5M
RP-ANN	361 (1.00)	680 (1.00)	2040 (1.00)	3400 (1.00)
SVD-ANN	434 (1.20)	1034 (1.52)	3427 (1.68)	6290 (1.85)
CRAFT-DP	384 (1.06)	762 (1.12)	2346 (1.15)	4012 (1.18)
CRAFT	415 (1.15)	823 (1.21)	2489 (1.22)	4182 (1.23)

6.4 Ablation and Comparative Analysis

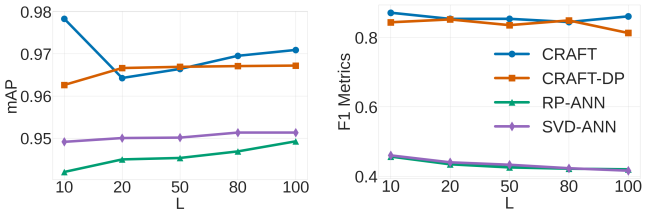
We conduct an ablation study to analyze the individual contributions and computational cost of CRAFT's key components, as well as to evaluate its sensitivity to its primary hyperparameters: the embedding dimension k, the nearest neighbor list size L, and the significance level α. This study uses a subsample of the **GenWiki** dataset with 100K documents. Unless otherwise noted, the default parameters are an embedding dimension of $k = 256$, a neighbor set size of $L = 100$, and a significance level $\alpha = 0.05$.



(a) mAP vs. k **(b) F1-Score vs. k**

Figure 3: Ablation study on the impact of embedding dimension (k) on performance.

6.4.1 Sensitivity to Embedding Dimension k. As shown in Figure 3a, mAP for both CRAFT variants peaks at $k = 256$, matching the approximate intrinsic dimensionality of the GenWiki semantic space; larger k introduces redundancy and a slight decline. The F1-Score (Figure 3b) plateaus for all $k \geq 256$, indicating that once the embedding captures the core relationships, performance is insensitive to further increases—a practical advantage that simplifies tuning.

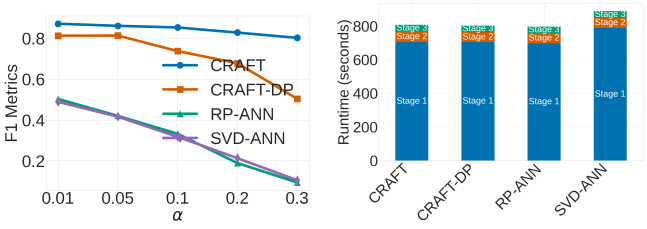


(a) mAP vs. L **(b) F1-Score vs. L**

Figure 4: Ablation study on the impact of nearest neighbor list size (L) on performance.

6.4.2 Sensitivity to Nearest Neighbor List Size L. Figure 4a shows that baseline mAP improves slightly with L (larger recall pool), while CRAFT's mAP is less dependent on list size because OMP seeks an optimal sparse representation rather than ranking by raw dot product. Both CRAFT variants maintain superior mAP across all L.

The F1-Score (Figure 4b) reveals a sharper distinction: for the baselines, F1 *decreases* with L as additional borderline candidates inflate false positives, indicating a poorly separated similarity measure. In contrast, CRAFT and CRAFT-DP remain stable and high, demonstrating a discriminative score distribution that reliably identifies significant neighbors regardless of list size.



(a) F1-Score vs. significance level (α). **(b) Runtime breakdown per stage.**

Figure 5: Ablation study on the impact of significance level (α) and a component-wise runtime analysis.

6.4.3 Sensitivity to Significance Level α. Figure 5a shows that F1 decreases with increasing α for all methods, as a looser threshold admits more false positives. **CRAFT** exhibits the highest F1 and the greatest stability across all α values, followed by CRAFT-DP; both substantially outperform the baselines. The robustness of CRAFT as α loosens confirms that its OMP-based scores are more discriminative, while SVD-ANN and RP-ANN degrade rapidly due to poorly separated similarity distributions.

6.4.4 Component-wise Runtime Analysis. Figure 5b breaks down runtime by stage on GenWiki. Stage 1 (Embedding Generation) dominates for all methods. SVD-ANN is the slowest overall due to its costly matrix factorization, while CRAFT, CRAFT-DP, and RP-ANN have comparable totals. Crucially, CRAFT's OMP-based Stage 3 adds negligible overhead relative to the simpler dot-product reranking, demonstrating that CRAFT's effectiveness gains come at virtually no additional computational cost.

6.4.5 Embedding Methods. To empirically distinguish semantic similarity from statistical correlation, we benchmark Model2Vec [49] against the Pearson correlation ground truth on the Wikidata dataset. By substituting the signal-generation stage with pre-trained vectors while keeping the standard retrieval protocol, we observed a marked performance gap. Model2Vec yielded a negligible F1 of 0.2 and mAP of 6.4, whereas CRAFT achieved 80.5 and 98.7 respectively. This gap highlights a fundamental objective mismatch. Embedding models optimize for predictive likelihood—grouping terms that constitute valid linguistic substitutions. While this captures broad semantic relatedness, it fails to preserve the precise, global linear covariation required for rigorous statistical discovery. The near-random performance of embeddings on the Pearson benchmark demonstrates that semantic proximity does not imply—and often obscures—statistical correlation.

6.4.6 Input Weighting and Sparsity Validation.

Sensitivity to input weighting. CRAFT's pipeline—Fourier embedding, ANN candidate filtering, and OMP sparse recovery—operates on any real-valued term-document matrix. By default we use BM25 weighting, but the choice of input weighting is orthogonal to the pipeline itself. To verify this, we replace the BM25 matrix with a Positive Pointwise Mutual Information (PPMI) matrix and re-run the full CRAFT pipeline on the GenWiki subsample used throughout this ablation study. Table 3 reports the results.

Table 3: Effect of input weighting on CRAFT (GenWiki).

Input Weighting	mAP	F1
BM25	85.0	81.3
PPMI	84.3	80.7

The difference is less than one percentage point on both metrics, confirming that CRAFT's effectiveness derives from the Fourier embedding and sparse recovery stages, not from properties specific to BM25.

Empirical sparsity and OMP convergence. CRAFT's compressed-sensing recovery assumes that each term's correlation vector is approximately sparse. We verify this across all datasets by computing brute-force pairwise Pearson correlations on vocabulary subsets and counting the number of significant correlations per term after Benjamini–Hochberg correction at $\alpha = 0.05$. Table 4 reports both the per-dataset sparsity distribution and the sensitivity of CRAFT to the OMP sparsity parameter S, which is the maximum number of nonzero entries OMP is allowed to select per query term.

Table 4: Sparsity validation. *Top:* distribution of significant correlations per term across all datasets, computed via brute-force on vocabulary subsets. *Bottom:* CRAFT performance as a function of the OMP sparsity budget S on the GenWiki.

(a) Significant correlations per term				
Dataset	**Mean**	**Median**	**99th %ile**	**Max**
Wikidata	1.1	1	3	4
DBpedia	1.0	1	2	4
GenWiki	1.3	1	3	6
arXiv-1M	1.4	1	3	7
arXiv-3M	1.5	2	4	8
arXiv-5M	1.5	2	4	9
(b) OMP sensitivity to sparsity parameter S				
S	5	10	100	200
mAP	84.8	85.0	85.0	84.9
F1	81.1	81.3	81.3	81.2

The correlation structure is consistently sparse across all datasets: the median term has 1–2 significant correlates, and even at the 99th percentile the count never exceeds 4. The trend is stable as corpus size grows—arXiv-5M, with over 5 million documents, shows a median of only 2, with a modest increase in the tail relative to the smaller benchmarks. Consistent with this strong sparsity, OMP's output on the GenWiki is effectively constant for all $S \geq 5$, with mAP and F1 varying by at most 0.2 points. OMP terminates well before reaching the S budget for the vast majority of terms, so S is not a sensitive hyperparameter in practice.

6.5 External Validation: Biomedical Case Study

The preceding experiments evaluate CRAFT on structured benchmarks whose ground truth is derived from correlation thresholds. To test whether CRAFT generalizes to noisy, heterogeneous real-world text and produces outputs that are meaningful under an independently curated ground truth, we collaborate with researchers from the biomedical domain to conduct an end-to-end case study [21]. This study applies CRAFT to large-scale biomedical literature and validates the resulting knowledge graph through both external ground-truth comparison and a downstream Retrieval-Augmented Generation (RAG) application.

6.5.1 Dataset and Domain Adaptation. We apply CRAFT to a corpus of 750,000 full-text PubMed papers from which 44.7 million text spans are extracted. This corpus poses three challenges absent from the benchmarks above: (i) *nomenclature ambiguity*—a single gene may appear under four or more aliases (e.g., ERBB2, HER2, NEU, CD340); (ii) *extreme noise*—raw spans mix experimental boilerplate (e.g., "Western blot", "incubated for 24 hours") with functional assertions; and (iii) *scale*—at 44.7 million spans, the corpus is nearly an order of magnitude larger than arXiv-5M.

Two domain-specific adaptations are required. First, we perform vocabulary normalization using the NCBI gene standardization protocol [32], mapping 262,097 known aliases to their canonical gene symbols so that, for example, mentions of "HER2" and "ERBB2" across different papers contribute to the same signal. Second, we construct *synthetic documents*: for each paper, all normalized gene mentions are aggregated into a single token sequence so that BM25 weighting operates over a vocabulary of gene symbols exclusively. These adaptations are preprocessing steps; the core CRAFT pipeline (Fourier embedding, HNSW candidate filtering, OMP sparse recovery) is applied without modification, producing a knowledge graph of 125,656 gene–gene pairs from 548,576 papers.

6.5.2 Independent Ground Truth. To avoid the circularity inherent in correlation-based benchmarks, we validate the extracted graph against the Kyoto Encyclopedia of Genes and Genomes (KEGG) [28]. KEGG comprises 582 pathway maps whose edges represent functional gene–gene relationships *manually curated by domain biologists* from experimental evidence—entirely independent of any co-occurrence statistic. From the aggregate pool of KEGG pathway edges, we randomly sample 1,000 interactions, each labeled as *activation* or *inhibition*, to form the evaluation set.

6.5.3 Downstream Task: RAG-Based Interaction Classification. We integrate the CRAFT-extracted graph into a RAG pipeline to test whether the graph provides useful context for a downstream reasoning task. For each query gene pair (A, B), the system retrieves text spans from the local graph neighborhoods of both A and B via neighbor sampling. These spans are concatenated and provided as context to a Large Language Model (LLM), which classifies the interaction as activation or inhibition. Crucially, the LLM must infer the relationship from the *functional community* surrounding each gene, even when the specific edge $A \to B$ is absent from the graph. We compare two conditions:

(1) **Baseline (Zero-Shot):** The LLM receives only the gene names and pathway identifier.
(2) **CRAFT-Augmented RAG:** The LLM receives the same prompt prepended with graph-retrieved context spans.

Table 5: Gene interaction classification on 1,000 KEGG edges.

Method	Accuracy	Balanced Acc.	F1
Zero-Shot LLM	0.77	0.77	0.71
LLM + CRAFT RAG	0.78	0.78	0.72

6.5.4 Results. Table 5 reports the results. The CRAFT-augmented RAG system improves accuracy and balanced accuracy from 0.77 to 0.78, and F1 from 0.71 to 0.72. The zero-shot baseline draws on knowledge absorbed during LLM pre-training over vast web-scale corpora—effectively representing the aggregate of publicly available human knowledge about gene interactions. CRAFT, a fully automated unsupervised pipeline operating solely on raw PubMed text, matches and exceeds this baseline on a ground truth curated independently by domain biologists, demonstrating that its statistical signal tracks genuine biological relationships rather than co-occurrence artifacts—a capability especially valuable in novel scientific domains and proprietary clinical repositories where curated KGs do not yet exist.

The qualitative impact is equally pronounced. The zero-shot baseline produces reasoning traces grounded in parametric memorization—assertions such as "CISH is annotated as a negative regulator" that are opaque and unverifiable [21]. The CRAFT-augmented model instead anchors its predictions in retrieved evidence (e.g., "The provided contexts describe SOCS/CIS proteins as suppressors of cytokine signaling…"), shifting from memorized assertions to evidence-grounded predictions. In scientific applications, this auditability is as valuable as the numerical improvement.

7 RELATED WORK

The fundamental task of identifying meaningful relationships between terms in a large corpus has been a long-standing challenge in natural language processing and information retrieval.

Statistical Association Measures. A foundational approach relies on computing statistical measures from a term co-occurrence matrix, such as Pointwise Mutual Information (PMI) [11] and the Chi-Squared (χ^2) test [34]. While intuitive for individual pairs, these methods suffer from inherent quadratic scaling; evaluating all pairs across a large vocabulary requires prohibitive $O(N^2)$ computations, a classic bottleneck in data mining [16, 53]. PMI also tends to overestimate the importance of rare term pairs, and χ^2 is difficult

to interpret as a metric for association strength. In contrast, the Pearson correlation coefficient offers a more suitable measure for identifying consistently covarying terms: it provides a normalized, bounded score in $[-1, 1]$ that captures both the strength and the direction of a linear relationship. Unlike PMI, Pearson correlation is less sensitive to sparse counts and effectively isolates proportional co-occurrence patterns, making it a robust choice for filtering significant entity pairs.

Neural and Embedding-Based Methods. Modern approaches leverage distributional semantics, learning dense vector representations of terms from their contexts. Models such as Word2Vec [36] and GloVe [41] infer relationships via cosine similarity in the embedding space, and pre-trained language models such as BERT [14] achieve strong performance on supervised relation-extraction tasks. However, training such embeddings or running inference with large models for every term pair is prohibitive for exhaustive discovery over massive corpora, and these methods conflate paradigmatic and syntagmatic relations. CRAFT avoids both issues by directly estimating the Pearson correlation coefficient.

Latent Factor and Topic Models. Latent Semantic Analysis (LSA) [13] applies Singular Value Decomposition (SVD) to a term-document matrix to project terms into a latent topic space. While effective, performing a full SVD on a large matrix is often too expensive for truly large-scale data. Moreover, the nature of the resulting associations differs fundamentally from ours: LSA relies on cosine similarity in a reduced space, which is strictly symmetric, lacks a universal scale, and reflects shared latent content. In contrast, Pearson correlation is bounded in $[-1, 1]$ and captures directional linear relationships in which one variable changes predictably with the other. CRAFT also bears some relationship to probabilistic topic models such as Latent Dirichlet Allocation (LDA), particularly in exploiting sparsity to discover latent structure. The objectives, however, are distinct: topic models optimize for broad semantic themes by clustering terms based on document-level co-occurrence probabilities, whereas CRAFT is specifically designed to estimate the Pearson correlation coefficient. This focus enables CRAFT to capture precise, directional linear dependencies (e.g., distinguishing positive from negative correlations), rather than the general thematic coherence in which such directional information is typically lost.

Random Projections. A highly scalable alternative is to use random projections, which preserve pairwise distances between points with high probability under the Johnson–Lindenstrauss lemma [25, 27]. The standard approach projects high-dimensional data onto a random lower-dimensional subspace. One could apply random projections directly to the term vectors derived from an inverted index; while feasible, this approach still requires associating N vectors, resulting in an $O(N^2)$ operation with significant overhead. Our method avoids this quadratic bottleneck entirely by operating directly on the compressed frequency-domain representation via sparse-recovery techniques. Furthermore, standard random projections are designed to preserve Euclidean geometry, whereas our Random Fourier Transform (RFT) is constructed specifically to produce a compact sketch that preserves the frequency-domain correlations and cross-power-spectral properties needed to estimate Pearson correlation. This provides a direct and computationally advantageous link between the time-domain signal (token occurrences) and the desired association metric.

Fourier Methods in Bioinformatics. Fourier transforms have a long history in computational biology, where character strings are mapped to numerical indicator sequences and analyzed spectrally. Voss [50] pioneered this line of work, converting DNA base sequences into binary indicator signals and revealing long-range $1/f$ fractal correlations via power-spectral analysis. Subsequent work applied the Discrete Fourier Transform (DFT) to detect periodicities in genomic sequences—most notably the three-base periodicity characteristic of protein-coding regions [47]—and to perform sequence alignment by identifying homologous regions through cross-spectral similarity [4]. While these bioinformatics methods and CRAFT both operate in the frequency domain, the objectives and constructions differ fundamentally. DNA methods apply the DFT to a *single, long* character sequence to uncover periodic patterns or local alignments *within* that sequence. In contrast, CRAFT uses a *randomized* Fourier transform (RFT) to sketch the occurrence frequencies of terms *across* a collection of discrete documents, producing compact embeddings whose inner products approximate global Pearson correlations between term pairs. The signal model (one long sequence vs. a term-document matrix), the transform (deterministic DFT vs. randomized subsampled DFT), and the goal (periodicity detection vs. pairwise correlation estimation) are all distinct.

8 CONCLUSION

Discovering related terms in massive, unstructured text corpora is a critical bottleneck for applications such as knowledge-graph construction. We presented CRAFT, a framework that addresses this challenge by reframing term discovery as a signal-processing task. CRAFT leverages randomized Fourier embeddings to compress term occurrences with theoretical guarantees, adapts the Cross-Power Spectral Density (CPSD) for robust correlation estimation, and employs sparse recovery via Orthogonal Matching Pursuit to identify significant relationships without materializing the prohibitive pairwise matrix. Our experimental evaluation shows that CRAFT outperforms state-of-the-art baselines in both precision and scalability across benchmark and large-scale corpora, and our biomedical case study against the KEGG pathway database demonstrates that CRAFT can bootstrap high-quality knowledge graphs in novel domains where curated KGs are unavailable. By acting as an efficient and statistically rigorous first-pass filter, CRAFT provides a scalable foundation for downstream knowledge-discovery pipelines and for augmenting large language models with verifiable, corpus-grounded evidence.

REFERENCES

[1] 2026. CRAFT: Corpus Relatedness Analysis Using Fourier Transforms (Technical Report). https://github.com/kckevinchen/CRAFT. Includes detailed proofs of all theorems and lemmas.

[2] Hassan Abdallah, Béatrice Markhoff, and Arnaud Soulet. 2025. Ranking Indicator Discovery from Very Large Knowledge Graphs. *Proc. VLDB Endow.* 18, 4 (May 2025), 1183–1195. https://doi.org/10.14778/3717755.3717775

[3] Dimitris Achlioptas. 2003. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. System Sci.* 66, 4 (2003), 671–687. https://doi.org/10.1016/S0022-0000(03)00025-4 Special Issue on PODS 2001.

[4] Dimitris Anastassiou. 2001. Genomic signal processing. *IEEE Signal Processing Magazine* 18, 4 (2001), 8–20. https://doi.org/10.1109/79.939833

[5] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: a nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference* (Busan, Korea) *(ISWC'07/ASWC'07)*. Springer-Verlag, Berlin, Heidelberg, 722–735.

[6] Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. 2008. A Simple Proof of the Restricted Isometry Property for Random Matrices. *Constructive Approximation* 28 (12 2008), 253–263. https://doi.org/10.1007/s00365-007-9003-x

[7] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300. http://www.jstor.org/stable/2346101

[8] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 2787–2795.

[9] Emmanuel Candes, Justin Romberg, and Terence Tao. 2004. Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information. arXiv:math/0409186 [math.NA] https://arxiv.org/abs/math/0409186

[10] Claudio Carpineto and Giovanni Romano. 2012. A Survey of Automatic Query Expansion in Information Retrieval. *ACM Comput. Surv.* 44, 1, Article 1 (Jan. 2012), 50 pages. https://doi.org/10.1145/2071389.2071390

[11] Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16, 1 (1990), 22–29. https://aclanthology.org/J90-1003/

[12] Colin B. Clement, Matthew Bierbaum, Kevin P. O'Keeffe, and Alexander A. Alemi. 2019. On the Use of ArXiv as a Dataset. arXiv:1905.00075 [cs.IR] https://arxiv.org/abs/1905.00075

[13] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 6 (1990), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL] https://arxiv.org/abs/1810.04805

[15] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). arXiv:2401.08281 [cs.LG]

[16] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han. 2014. Scalable topical phrase mining from text corpora. *Proc. VLDB Endow.* 8, 3 (Nov. 2014), 305–316. https://doi.org/10.14778/2735508.2735519

[17] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. *From data mining to knowledge discovery: an overview*. American Association for Artificial Intelligence, USA, 1–34.

[18] Simon Foucart and Holger Rauhut. 2013. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel.

[19] Cong Fu, Chao Xiang, Changxu Wang, and Deng Cai. 2019. Fast approximate nearest neighbor search with the navigating spreading-out graph. *Proc. VLDB Endow.* 12, 5 (Jan. 2019), 461–474. https://doi.org/10.14778/3303753.3303754

[20] Leslie Greengard and June-Yub Lee. 2004. Accelerating the Nonuniform Fast Fourier Transform. *SIAM Rev.* 46, 3 (2004), 443–454. https://doi.org/10.1137/S003614450343200X arXiv:https://doi.org/10.1137/S003614450343200X

[21] Sidharth Gupta. 2025. Grounding LLM Gene Interaction Predictions using Sparse Knowledge Graphs by CRAFT. University of Toronto, technical report.

[22] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. 2010. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. arXiv:0909.4061 [math.NA] https://arxiv.org/abs/0909.4061

[23] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge Graphs. *Comput. Surveys* 54, 4 (July 2021), 1–37. https://doi.org/10.1145/3447772

[24] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *Comput. Surveys* 55, 12 (March 2023), 1–38. https://doi.org/10.1145/3571730

[25] Jiawei Jiang, Fangcheng Fu, Tong Yang, and Bin Cui. 2018. SketchML: Accelerating Distributed Machine Learning with Data Sketches. In *Proceedings of the 2018 International Conference on Management of Data* (Houston, TX, USA) *(SIGMOD '18)*. Association for Computing Machinery, New York, NY, USA, 1269–1284. https://doi.org/10.1145/3183713.3196894

[26] Zhijing Jin, Qipeng Guo, Xipeng Qiu, and Zheng Zhang. 2020. GenWiki: A Dataset of 1.3 Million Content-Sharing Text and Graphs for Unsupervised Graph-to-Text Generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 2398–2409. https://doi.org/10.18653/v1/2020.coling-main.217

[27] William B Johnson, Joram Lindenstrauss, et al. 1984. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics* 26, 189-206 (1984), 1.

[28] Minoru Kanehisa. 2002. The KEGG database. In *'In Silico' Simulation of Biological Processes: Novartis Foundation Symposium 247*. Vol. 247. Wiley, 91–103.

[29] Shantanu Kumar. 2017. A Survey of Deep Learning Methods for Relation Extraction. arXiv:1705.03645 [cs.CL] https://arxiv.org/abs/1705.03645

[30] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.

[31] Kejing Lu, Mineichi Kudo, Chuan Xiao, and Yoshiharu Ishikawa. 2021. HVS: hierarchical graph structure based on voronoi diagrams for solving approximate nearest neighbor search. *Proc. VLDB Endow.* 15, 2 (Oct. 2021), 246–258. https://doi.org/10.14778/3489496.3489506

[32] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. 2007. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Research* 35, suppl_1 (2007), D26–D31. https://doi.org/10.1093/nar/gkl993

[33] Yu A. Malkov and D. A. Yashunin. 2020. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 4 (April 2020), 824–836. https://doi.org/10.1109/TPAMI.2018.2889473

[34] Christopher D Manning and Hinrich Schütze. 1999. Foundations of statistical natural language processing.

[35] Stefano Marchesin and Gianmaria Silvello. 2025. Credible Intervals for Knowledge Graph Accuracy Estimation. *Proc. ACM Manag. Data* 3, 3, Article 142 (June 2025), 26 pages. https://doi.org/10.1145/3725279

[36] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. http://arxiv.org/abs/1301.3781

[37] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li (Eds.). Association for Computational Linguistics, Suntec, Singapore, 1003–1011. https://aclanthology.org/P09-1113/

[38] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel. 2011. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on International Conference on Machine Learning* (Bellevue, Washington, USA) *(ICML'11)*. Omnipress, Madison, WI, USA, 809–816.

[39] Reham Omar, Omij Mangukiya, and Essam Mansour. 2025. Dialogue Benchmark Generation from Knowledge Graphs with Cost-Effective Retrieval-Augmented LLMs. *Proc. ACM Manag. Data* 3, 1, Article 31 (Feb. 2025), 26 pages. https://doi.org/10.1145/3709681

[40] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2024. Unifying Large Language Models and Knowledge Graphs: A Roadmap. *IEEE Transactions on Knowledge and Data Engineering* 36, 7 (July 2024), 3580–3599. https://doi.org/10.1109/tkde.2024.3352100

[41] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). Association for Computational Linguistics, Doha, Qatar, 1532–1543. https://doi.org/10.3115/v1/D14-1162

[42] Ali Rahimi and Benjamin Recht. 2007. Random Features for Large-Scale Kernel Machines. In *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis (Eds.), Vol. 20. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2007/file/013a006f03dbc5392effeb8f18fda755-Paper.pdf

[43] Diego Rincon-Yanez and Sabrina Senatore. 2022. FAIR Knowledge Graph construction from text, an approach applied to fictional novels. In *Proceedings of*

the 1st International Workshop on Knowledge Graph Generation From Text and the 1st International Workshop on Modular Knowledge co-located with 19th Extended Semantic Conference (ESWC 2022). CEUR-WS, Hersonissos, Greece, 94–108. http://ceur-ws.org/Vol-3184/TEXT2KG_Paper_7.pdf

[44] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (April 2009), 333–389. https://doi.org/10.1561/1500000019

[45] Florin Rusu and Alin Dobra. 2007. Statistical analysis of sketch estimators. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data* (Beijing, China) *(SIGMOD '07)*. Association for Computing Machinery, New York, NY, USA, 187–198. https://doi.org/10.1145/1247480.1247503

[46] Chuan Shi, Yitong Li, Jiawei Zhang, Yizhou Sun, and Philip S. Yu. 2017. A Survey of Heterogeneous Information Network Analysis. *IEEE Trans. on Knowl. and Data Eng.* 29, 1 (Jan. 2017), 17–37. https://doi.org/10.1109/TKDE.2016.2598561

[47] Shrish Tiwari, S. Ramachandran, Alok Bhattacharya, Sudha Bhattacharya, and Ramakrishna Ramaswamy. 1997. Prediction of probable genes by Fourier analysis of genomic sequences. *Bioinformatics* 13, 3 (1997), 263–270. https://doi.org/10.1093/bioinformatics/13.3.263

[48] Joel A. Tropp and Anna C. Gilbert. 2007. Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit. *IEEE Transactions on Information Theory* 53, 12 (2007), 4655–4666. https://doi.org/10.1109/TIT.2007.909108

[49] Stephan Tulkens and Thomas van Dongen. 2024. *Model2Vec: Fast State-of-the-Art Static Embeddings*. https://doi.org/10.5281/zenodo.17270888

[50] Richard F. Voss. 1992. Evolution of long-range fractal correlations and 1/f noise in DNA base sequences. *Physical Review Letters* 68, 25 (1992), 3805–3808. https://doi.org/10.1103/PhysRevLett.68.3805

[51] Feiyu Wang, Qizhi Chen, Yuanpeng Li, Tong Yang, Yaofeng Tu, Lian Yu, and Bin Cui. 2023. JoinSketch: A Sketch Algorithm for Accurate and Unbiased Inner-Product Estimation. *Proc. ACM Manag. Data* 1, 1, Article 81 (May 2023), 26 pages. https://doi.org/10.1145/3588935

[52] Mengzhao Wang, Xiaoliang Xu, Qiang Yue, and Yuxiang Wang. 2021. A comprehensive survey and experimental comparison of graph-based approximate nearest neighbor search. *Proc. VLDB Endow.* 14, 11 (July 2021), 1964–1978. https://doi.org/10.14778/3476249.3476255

[53] Hao Yan, Shuming Shi, Fan Zhang, Torsten Suel, and Ji-Rong Wen. 2010. Efficient term proximity search with term-pair indexes. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (Toronto, ON, Canada) *(CIKM '10)*. Association for Computing Machinery, New York, NY, USA, 1229–1238. https://doi.org/10.1145/1871437.1871593

[54] Huiyuan Yu, Jia He, and Maggie Cheng. 2025. Fast Orthogonal Matching Pursuit through Successive Regression. arXiv:2404.00146 [cs.CV] https://arxiv.org/abs/2404.00146

[55] Xi Zhao, Yao Tian, Kai Huang, Bolong Zheng, and Xiaofang Zhou. 2023. Towards Efficient Index Construction and Approximate Nearest Neighbor Search in High-Dimensional Spaces. *Proc. VLDB Endow.* 16, 8 (April 2023), 1979–1991. https://doi.org/10.14778/3594512.3594527