# Multi-Vector Models with Textual Guidance for Fine-Grained Scientific Document Similarity

**Anonymous ACL submission**


## Abstract

We present a new scientific document similarity model based on matching fine-grained aspects of texts. To train our model, we exploit a naturally-occurring source of supervision: sentences in the full-text of papers that cite multiple papers together (*co-citations*). Such co-citations not only reflect close paper relatedness, but also provide textual descriptions of *how* the co-cited papers are related. This novel form of *textual supervision* is used for learning to match aspects across papers. We develop multi-vector representations where vectors correspond to sentence-level aspects of documents, and present two methods for aspect matching: (1) A fast method that only matches single aspects, and (2) a method that makes sparse multiple matches with an Optimal Transport mechanism that computes an Earth Mover's Distance between aspects. Our approach improves performance on document similarity tasks in four datasets. Further, our fast single-match method achieves competitive results, paving the way for applying fine-grained similarity to large scientific corpora.[1]


## 1 Introduction

The ability to identify similarity across documents in large scientific corpora is fundamental for many applications, including recommendation (Bhagavatula et al., 2018), exploratory or analogical search (Hope et al., 2017, 2021b; Lissandrini et al., 2019), paper-reviewer matching (Mimno and McCallum, 2007; Berger et al., 2020) and many more uses.

Scientific papers often describe multifaceted arguments and ideas (Hope et al., 2021a; Lahav et al., 2021), suggesting that models capable of matching specific aspects can better capture overall document relatedness, too. For example, sentences in research abstracts can often be categorized as descriptions of objectives, methods, or findings (Kim et al., 2011; Chan et al., 2018), centrally important discourse structures of scientific texts.

In this paper, we propose a new model for document similarity that makes aspect-level matches across papers and aggregates them into a document-level similarity. We focus on sentence-level aspects of paper abstracts, and train multi-vector representations of papers in terms of their contextualized sentence embeddings. To train our models, we leverage a readily available data source: sentences that co-cite multiple papers. Unlike recent work that used citation links for learning scientific document similarity (Cohan et al., 2020), we observe that papers cited in close proximity provide a more precise indication of relatedness. Furthermore, the citing sentences typically describe how the co-cited papers are related, in terms of shared aspects (e.g., similar methods or findings, related challenges or directions, etc.). Building on this observation, we leverage these textual descriptions as a novel source of *textual supervision*, using them to guide our model to learn which sentence-aspects match without any direct sentence-level supervision. Guidance for the document similarity model is obtained via an auxiliary sentence encoder model that is used for aligning abstract sentences by finding pairs most similar to the citing sentence text.

Our document similarity objective is modeled as a function of similarity between sentence-level matches. We explore two strategies to aggregate over sentence-level distances between documents. First, a single-match method with minimum L2 distances between document aspect vectors. This approach readily supports approximate nearest neighbor search methods for large-scale retrieval. Second, a multi-match method that computes an Earth Mover's Distance between documents' aspect vectors by solving an Optimal Transport problem. This yields a soft sparse matching of aspect vectors, which when combined with their L2 distances gives a document-level distance.

[1] Code available at: `https://anonymous.4open.science/r/aspire-F570`

Finally, as an additional benefit of our representation, our models also support a finer *aspect-conditional* retrieval task (Hope et al., 2017, 2021a; Chan et al., 2018; Mysore et al., 2021) where aspects can be specified by selecting abstract sentences — for example, selecting sentences describing methods and retrieving papers using similar methods. As we show, naively encoding sentences without their context leads to subpar results in this task, and our representation that does take context into account dramatically improves results.

Extensive empirical evaluation on four English scientific text datasets and seven similarity tasks at the level of documents and sentences demonstrates the effectiveness of our models. These include biomedical document retrieval tasks and a recent faceted query-by-example corpus of computer science papers (Mysore et al., 2021). This latter dataset is used for evaluating retrieval conditioned on specific aspects in context (e.g., for finding papers with similar methods to a query document), demonstrating that our model can be used in this challenging and important setting. In summary, we make the following main contributions:

1. **Multi-Vector Document Similarity Model**: We present ASPIRE[2], a multi-vector document similarity model that flexibly aggregates over fine-grained sentence-level aspect matches.
2. **Co-Citation Context Supervision**: We exploit widely-available co-citation sentences as a new source of training data for document similarity and provide a method using a novel form of textual supervision to guide representation learning for aspect matching.
3. **State of the Art Results**: Our ASPIRE models outperforms strong baseline methods across four datasets for the abstract and aspect-conditional similarity tasks.

## 2 Problem Setup

Given query document $Q$ and a candidate amongst a set of documents $C \in \mathcal{C}$, where documents consists of $N$ sentences $\langle S_1, S_2, \dots S_N \rangle$ we aim to leverage fine-grained document similarity in two problem settings. An abstract level retrieval task (Brown et al., 2019; Cohan et al., 2020) and an aspect-level retrieval task (Mysore et al., 2021):

**Def 1.** *Retrieval by abstracts: Given query and candidate documents – $Q$ and $\mathcal{C}$ a system must output the ranking over $\mathcal{C}$.*

[2]ASPIRE: Aspectual Scientific Paper Relations.

**Def 2.** *Aspect-level retrieval by sentences: Given query and candidate documents – $Q$ and $\mathcal{C}$, and a subset of sentences $\mathcal{S}_Q \subseteq Q$ conditional on which to retrieve documents, a system must output the ranking over $\mathcal{C}$.*

**Modeling Desiderata:** Next, we also outline key desired properties we require from models developed for task definitions 1 and 2. We follow these desiderata when building our methods (§3.1).

1. *Allowing specification of optional fine-grained aspects*: We would like models to allow the ability to specify fine-grained query aspects in a query document based on which retrievals should be made. These may be obtained automatically (e.g., with a discourse tagging method) or via user specification.
2. *Scalable to large corpora and efficient inference*: State of the art retrieval systems often rely on expensive cross-attention mechanisms on query-document pairs making training and inference expensive (Zamani et al., 2018; Lin et al., 2021). This is exacerbated for longer scientific documents requiring specific transformer models (Caciularu et al., 2021). We require our methods to leverage large training corpora and allow efficient inference at scale.

## 3 Proposed Approach: ASPIRE

In this section we describe our approach to document similarity – ASPIRE. We model finer-grained matches between documents at the level of sentences via contextual representations and aggregating over matches to obtain similarities between whole documents. We leverage co-citation sentences as a source of document similarity and also as implicit *textual supervision* describing related aspects of co-cited documents. We formulate our multi-vector models (Luan et al., 2021; Humeau et al., 2020) that can support scalable inference as novel multiple-instance learning (MIL) models.

### 3.1 Fine-grained Document Similarity

We assume to be given a training set consisting of sets of documents $\mathcal{P}$ which are *weakly-labeled* for similarity. We leverage widely-available sets of papers co-cited together in the same sentence as similar (see Figure 1). This builds on the observation that co-citations in close proximity (e.g., in the same sentence) are strong indicators for paper relatedness (Gipp and Beel, 2009).

We follow the contrastive learning framework, commonly used for learning semantic similarity (Reimers and Gurevych, 2019; Cohan et al., 2020).

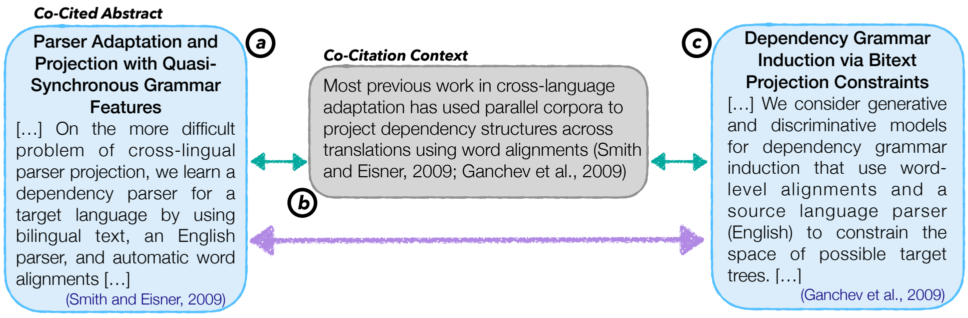



Figure 1: Example illustrating the signal in co-citations. Of all the sentences in co-cited abstracts ⓐ and ⓒ the sentences shown are each individually aligned to co-citation context ⓑ as per embeddings from $\texttt{BERT}_{\mathcal{E}}$ (§3.2.2). Consequently these sentences in ⓐ and ⓒ are treated as sharing aspects between the co-cited papers to supervise our fine-grained similarity model for single matching.

We train models on triples of the form $(p, p', n)$ where $p, p' \in \mathcal{P}$ and $n \notin \mathcal{P}$ is a randomly selected negative, using the triple margin ranking loss $\mathcal{L}_f(p, p', n) = \texttt{max}[f(p, p') - f(p, n) + m, 0]$, where $f(\cdot, \cdot)$ is a distance between documents. All pairwise-combinations $p, p' \in \mathcal{P}$ are treated as positive pairs in-turn. In this work, we parameterize $f$ based on the distances between finer-grained document aspects $\mathcal{A}$. Given documents $p$ and $p'$, we focus on a family of functions $f$ of the form:

$$f(p, p') = \sum_{(i,i') \in \mathcal{A}_p \times \mathcal{A}_{p'}} w_{i,i'} \cdot d_{i,i'}. \qquad (1)$$

Here, $\mathcal{A}_p \times \mathcal{A}'_p$ represents the space of alignments between aspects of document $p$ and $p'$, $d_{i,i'}$ denotes a distance between two aspects $i, i'$, and $w_{i,i'}$ represents a weight indicating the contribution of the aspect similarity to the overall document similarity. Unlike previous work (Neves et al., 2019; Jain et al., 2018; Hope et al., 2017), we make no assumption on specific aspect semantics in deriving a model architecture, and focus on aspects in the form of *general* subsets of document sentences.

For learning, we only assume to be given *document-level* supervision (sets of documents $\mathcal{P}$), and no supervision on *aspect-level* similarity. Our task thus consists of learning $w_{i,i'}$ and $d_{i,i'}$ via indirect supervision. We cast this problem setting as a novel type of multi-instance learning (MIL) (Ilse et al., 2018) problem. Prior work in MIL broadly aims to learn instance level classifiers given labels for a bag of instances, this bears resemblance to our setting, where instances are aspects $\mathcal{A}$. However, unlike prior MIL work we focus on learning *similarity* rather than *classification*. We formulate two variants of $f$ in Equation 1:

(1) A single match model (§3.2.2) which considers documents similar based on the single most similar alignment $\hat{i}_p, \hat{i}_{p'} \in \mathcal{A}_p \times \mathcal{A}'_p$. This assumes $w = 1$ for the best alignment and $w = 0$ elsewhere.

(2) A multi match model (§3.2.3) which makes multiple alignments between documents. We find aspect importance weights $w_{i,i'}$, by solving an Optimal Transport (OT) problem (Peyré et al., 2019).

In both variants, during training we learn contextualized aspect embeddings that minimize the contrastive loss parametrized with $f$.

**Co-citation Contexts as Supervision:** Finally, we present a method for incorporating implicit natural language supervision during training, presented by co-citation sentences which describe specific relations between co-cited documents. For example, Figure 1 shows a case explaining the similarity between the co-cited papers' methods. We leverage this textual supervision to find a "best" alignment $\hat{i}_p, \hat{i}_{p'}$ in the single-alignment variant (1), and for guiding the optimal transport plan in variant (2). We describe the specific model components next.

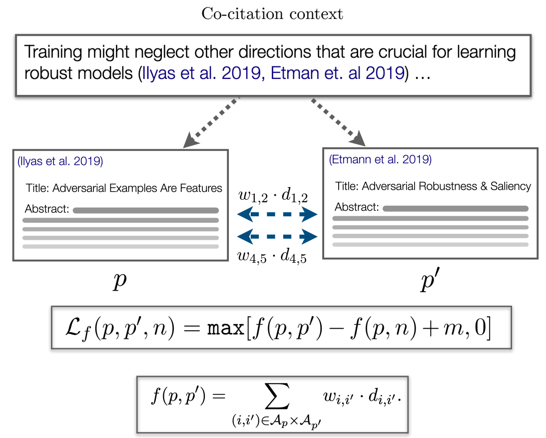
## 3.2 Model Description

### 3.2.1 Document Encoder

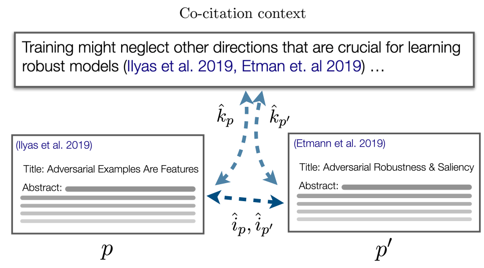
We leverage a pre-trained BERT-based language model as a document encoder as the base of all our methods. Our encoder is mainly intended to output contextualized sentence representations. Given a document title and abstract, this is achieved as:

$$\mathbf{S} = \texttt{BERT}_\theta([\texttt{CLS}]\ \texttt{Title}\ [\texttt{SEP}]\ \texttt{Abstract}) \qquad (2)$$

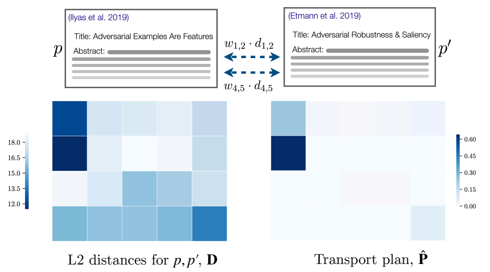
where $\mathbf{S} \in \mathbb{R}^{N \times d}$ represents contextualized sentences $\mathbf{s}_1 \ldots \mathbf{s}_N$ stacked into a matrix. Here, each $\mathbf{s}$ is obtained by mean-pooling word-piece embeddings from the final layer of $\texttt{BERT}_\theta$ for the sentence tokens. Pairwise distances between sentences




(a) Learning fine-grained document similarity using co-citations.




(b) The single best match, $\hat{i}_p, \hat{i}_{p'}$, is computed from textual supervision in the co-citation context.




(c) Multi-aspect matches via a sparse transport plan.

Figure 2: Approach overview. (a) We train fine-grained similarity models using papers co-cited in the the same sentence in research papers. (b) Single-match models are learned from implicit supervision in co-citation contexts. (c) Multi-match models are learned by aligning aspect representations by solving an Optimal Transport problem.

$d_{i,i'}$ in Eq 1 for $p, p'$, are represented as a matrix $\mathbf{D} \in \mathbb{R}^{N \times N'}$ of L2 distances between $\mathbf{S}_p$ and $\mathbf{S}_{p'}$.

#### 3.2.2 Single Match & Textual Supervision

Our single match model makes the assumption that document similarity is explained by a single best match, giving $f_{\mathtt{TS}}(p, p') = \mathbf{D}[\hat{i}_p, \hat{i}_{p'}]$. Here, we leverage weak supervision from co-citation contexts for training. This is done by using an auxiliary sentence encoder to compute a maximally aligned sentence $\hat{i}_p$ in co-cited paper $p$ to the co-citation context, similarly $\hat{i}_{p'}$ aligns a sentence in $p'$ to the co-cotation context. Then the two context aligned sentences are treated as aligned to each other, for training. In practice, the same papers $\mathcal{P}$ can be co-cited in multiple different papers (in $\sim 30\%$ of co-cited papers) giving us a set of co-citation sentences, $e \in \mathcal{E}$ and training data of the form $(\mathcal{E}, \mathcal{P})$. Alignments of the sentences in $p$ and $p'$ to the co-citation contexts $e \in \mathcal{E}$ are computed as:

$$\begin{aligned} \hat{i}_p, \hat{k}_p &= \operatorname*{argmax}_{i=1\ldots N, k=1\ldots N'} \mathbf{R}_p \mathbf{R}_{\mathcal{E}}^T \\ \hat{i}_{p'}, \hat{k}_{p'} &= \operatorname*{argmax}_{i=1\ldots N, k=1\ldots N'} \mathbf{R}_{p'} \mathbf{R}_{\mathcal{E}}^T \end{aligned} \tag{3}$$

Here $\mathbf{R}_p$, $\mathbf{R}_{p'}$, and $\mathbf{R}_{\mathcal{E}}^T$ are independent sentence representations for $p, p'$ and $e$, respectively, obtained from a auxiliary sentence encoder $\mathtt{BERT}_{\mathcal{E}}$ (details below), and $\hat{i}_p, \hat{i}_{p'}$ represent the single best alignment of sentences across $p, p'$ "anchored" on textual supervision sentences $\mathcal{E}$. Importantly, this supervision is only used during training time to guide learning. This procedure is depicted in Figure 2b with Figure 1 showing an example.

**Co-citation Context Encoder** The encoder $\mathtt{BERT}_{\mathcal{E}}$ represents a SciBERT based sentence encoder pre-trained for scientific text similarity. We train $\mathtt{BERT}_{\mathcal{E}}$ on sets of co-citation contexts referencing the same set of papers (i.e. $\mathcal{E}$) in a contrastive learning setup. This set, $\mathcal{E}$, can be considered as paraphrases since co-citation sentences citing the same papers often describe similar relations between the papers This model is similar to Sentence-BERT (Reimers and Gurevych, 2019) and we refer to it as `CoSentBert`. In training document encoder $\mathtt{BERT}_\theta$, we keep $\mathtt{BERT}_{\mathcal{E}}$ frozen. Appendix C presents more detail on $\mathtt{BERT}_{\mathcal{E}}$ design.

#### 3.2.3 Multiple Matches & Optimal Transport

While a single best sentence alignment $\hat{i}_p, \hat{i}_{p'}$ may sufficiently explain document similarity for some documents and applications, documents often have a stronger and weaker alignments. So, in computing sentence alignments between documents we would like a sparse matching that aptly weights alignments while ignoring non-alignments — corresponding to learning weights $w_{i,i'}$ in Eq 1. To model this intuition we leverage optimal transport.

**Optimal Transport** The OT problem is constituted by two sets of points, $\mathbf{S}_p$ and $\mathbf{S}_{p'}$ as in our case, and distributions $\mathbf{x}_p$ and $\mathbf{x}_{p'}$ according to which the set of points is distributed. The OT problem involves computation of a transport plan $\hat{\mathbf{P}}$, which converts $\mathbf{x}_p$ into $\mathbf{x}_{p'}$ by transporting probability mass while minimizing an aggregate cost computed from the pairwise costs $\mathbf{D}$ of aligning the points in $\mathbf{S}_p$ and $\mathbf{S}_{p'}$. $\hat{\mathbf{P}}$ is constrained such that its columns and rows marginalize respectively to $\mathbf{x}_p$ and $\mathbf{x}_{p'}$ (so that all mass is accounted for). Specifically, the computation of $\hat{\mathbf{P}}$ takes the form of a constrained linear optimization problem:

$$\mathcal{W} = \min_{\mathbf{P}\in\mathcal{S}} \langle \mathbf{D}, \mathbf{P} \rangle \tag{4}$$

$$= \min_{\mathbf{P}\in\mathcal{S}} \sum_{i=1}^{N} \sum_{j=1}^{N'} \mathbf{D}[i,j]\mathbf{P}[i,j] \tag{5}$$

$$\mathcal{S} = \{\mathbf{P} \in \mathbb{R}_{+}^{N\times N'} | \mathbf{P}\mathbf{1}_{N'} = \mathbf{x}_p, \mathbf{P}^T\mathbf{1}_{N'} = \mathbf{x}_{p'}\} \tag{6}$$

where $\mathcal{W}$ refers to the Wasserstein or Earth Movers Distance and $\hat{\mathbf{P}}$ is the minimizer resulting from solving Eq 5. Of interest here is an established result which shows $\hat{\mathbf{P}}$ to be sparse with $\mathcal{O}(N+N')$ non-zero entries (Swanson et al., 2020). Therefore, $\hat{\mathbf{P}}$ represents a soft sparse alignment of sentences and can be used as weights $w_{i,i'}$ in Eq 1, with document distances computed as $f_{\texttt{OT}}(p,p') = \langle \mathbf{D}, \hat{\mathbf{P}} \rangle$. Fig 2c presents a schematic for this approach.

Note that $\mathbf{x}_p$ and $\mathbf{x}_{p'}$ allow control over importance of sentences in $p$ and $p'$ in the form of relative probability mass. We compute these distributions using pairwise distances as $\mathbf{x} = \texttt{softmax}(-\mathbf{s}/\tau)$ where $\mathbf{s}_p = \texttt{min}_i\ \mathbf{D}$ and $\mathbf{s}_{p'} = \texttt{min}_j\ \mathbf{D}$, and $\tau$ is a softmax temperature hyper-parameter.

For our neural network models trained with automatic differentiation, we leverage an entropy regularized version of the Wasserstein distance in Eq 5 (Cuturi, 2013). Here computation of $\hat{\mathbf{P}}$, is achieved via Sinkhorn iterations, a set of iterative linear updates allowing training with autodiff libraries and leveraging GPU computation. Finally, Cuturi (2013) show that computing $\mathcal{W}$ with Sinkhorn iterations shows an empirical quadratic complexity, i.e. $\mathcal{O}(N^2)$ — similar to that of attention as in a model for late interaction (Humeau et al., 2020).

**Multi-task model:** To leverage training signals used in both the single and multi-match models, we train a multi match model supervised with textual supervision in a multi-task setup: $\mathcal{L}_{f_{\texttt{TS}}} + \mathcal{L}_{f_{\texttt{OT}}}$.

### 3.3 Inference

As outlined in §2, we are interested in a whole-abstract based retrieval (Def 1) and an aspect level retrieval (Def 2). In both setups given a query $Q$ and candidate $C$ documents we denote sentence representations from a trained model by $\mathbf{S}_Q$ and $\mathbf{S}_C$. For both tasks, we compute distances for ranking while controlling the aspects $\mathcal{A}_Q$ (i.e $\mathcal{A}_p$) over which the weighted sum of Eq 1 is performed.

**Whole abstract retrieval:** This corresponds to a setup where all aspects of the query document $\mathcal{A}_p$ are used in computing distances between documents. In the single-alignment models, candidates $C$ are ranked based on their maximally aligned sentence with $Q$ using distances from a trained model: $\hat{i}_p, \hat{i}_{p'} = \texttt{argmin}_{i,j}\ \mathbf{D}$. The multi match model ranks candidates using the distance $\langle \mathbf{D}, \hat{\mathbf{P}} \rangle$, where $\hat{\mathbf{P}}$ is the solution to transport problem of Eq 5.

**Aspect level retrieval:** In aspect-level retrieval, a subset of sentences $\mathcal{A}_q \subset \mathcal{A}_Q$ is used for query document $Q$; for candidate documents $C$, we do not assume to be given specific aspects, and matching is done across all sentences in each $C$. In the single alignment models, we only consider a subset of the pairwise sentence distances to determine the maximally aligned sentences, giving $\mathbf{D}^{\mathcal{A}} = \mathbf{D}[\mathcal{A}_q, :]$. This corresponds to finding the maximally aligned candidate sentence to the query sentences in $\mathcal{A}_q$. Similarly, in the multiple-alignment model we compute the plan $\hat{\mathbf{P}}^{\mathcal{A}_q}$ based on the subset of sentences corresponding to $\mathcal{A}_q$ and generate rankings by $\langle \mathbf{D}^{\mathcal{A}_q}, \hat{\mathbf{P}}^{\mathcal{A}_q} \rangle$. Note that $\mathbf{S}_Q$ in $Q$ is still contextualized, capturing document context of sentences not explicitly used in $\mathcal{A}_q$.

**Scaling Inference:** Our multi-vector model for single matching performs retrievals via minimum L2 distance. Therefore, this method is amenable to approximate nearest neighbour (ANN) search methods for large-scale retrieval (Andoni et al., 2018; Luan et al., 2021). Retrieval with our single-match model would involve $|\mathcal{A}_Q|$ and $|\mathcal{A}_q|$ calls to an ANN structure for the whole abstract and aspect-level tasks respectively.

On the other hand, as stated earlier our multi-match model using Sinkhorn iterations involves a $\mathcal{O}(N^2)$ computation (Cuturi, 2013), which is similar to late interaction methods. Humeau et al. (2020) show late interaction models to be significantly cheaper than cross-encoders while retaining most of their performance in ad-hoc search setups. While quadratic, OT computation in practice can be time-consuming, however, recent work of Backurs et al. (2020) has seen development of fast ANN methods for Wasserstein distances with practical run-times significantly smaller than quadratic ones. This promises the use of ANN methods in large-scale retrieval with our multi-match model

In our results we refer to our text supervised single match method as TSASPIRE, optimal transport multi match method as OTASPIRE, and the multi-task trained multi aspect method as TS+OTASPIRE.

## 4 Experiments and Results

**Evaluation data:** We evaluate the proposed methods on datasets for whole abstract document similarity and fine-grained document similarity. We overview these below. Appendix B provides detail.

1. RELISH: An expert annotated dataset of biomedical abstract similarity (Brown et al., 2019).

2. $\text{TRECCOVID}_{\text{RF}}$: The original TRECCOVID dataset is labelled for ad-hoc search by experts (Voorhees et al., 2021). We reformulate the dataset for abstract similarity, treating all abstracts relevant to one ad-hoc query as similar to each other and dissimilar from abstracts relevant to other queries.

3. SCIDOCS: A benchmark suite of tasks intended for evaluating abstract-level scientific document representations (Cohan et al., 2020).

4. CSFCUBE: Fine-grained retrieval is evaluated using the recent dataset of Mysore et al. (2021), an expert-annotated dataset of machine learning and NLP abstracts labelled against candidates for relevance to one of 3 broad aspects capturing the main components of methodological research: `background/objective`, `method`, `result`. Relevance is labelled for query sentences corresponding to those aspects, while considering the broader relevance of the sentences' abstract context.

**Baselines:** We compare the proposed approaches to three classes of methods. We overview these classes and associated models below, with Appendix D presenting further detail: 1. Sentence models: Sentence embedding models present reasonable baselines since we consider fine-grained matches at the sentence level. These are represented by MPNET-1B, a sentence model trained on over 1 billion text pairs[3], Sentence-Bert (SENTBERT) (Reimers and Gurevych, 2019), SIMCSE (Gao et al., 2021), `cosentbert` of §3.2.2, and ICTSENTBERT (Lee et al., 2019).

2. Abstract models: The abstract level model SPECTER (Cohan et al., 2020), represents a SOTA model for scientific document similarity trained on *cited* abstract pairs. We also train a variant of this model on *co-cited* papers: SPECTER-COCITE.

3. Sentence level models modified for whole abstract similarity: Here we combine the SOTA sentence encoder MPNET-1B with the optimal transport (§3.2.3) for aggregating sentence level matches giving OTMPNET-1B.

Sentence models use the same inference procedure as our single match method, abstract models rank using L2 distances between papers embeddings, and the modified sentence model uses the multi match inference procedure. All reported model hyper-parameters are tuned, trained on 1.3M co-citation triples, and initialized with SPECTER unless noted otherwise.[4] Appendices A, E, and F detail training data, algorithms, and hyper-parameters. Next, we present our main results comparing proposed approaches to baselines.

### 4.1 Results

**Fine-grained similarity**: Table 1 presents results on CSFCUBE. We report performance on the three facets `background`, `method`, and `result` annotated in the dataset, and aggregated across all facets. We first make some observations about baseline methods: 1. MPNET-1B outperforms all other sentence level models and a SOTA abstract representation, SPECTER, indicating the value of sentence-level information for capturing fine-grained similarities. With OTMPNET-1B indicating the value of modeling multiple matches. 2. $\text{SPECTER–COCITE}_{\text{Scib}}$, which is identical to SPECTER but trained on co-citations outperforms it, showing the value of co-citations for fine-grained similarity.

Next, we examine performance of the proposed methods: 1. First we note that all of the proposed approaches consistently outperform performant prior work, OT/MPNET-1B and SPECTER, by about 5-6 points aggregated across queries. 2. Next, we note that the proposed approaches outperform $\text{SPECTER-COCITE}_{\text{Spec}}$, trained on co-citations by 2-3 points aggregated across queries. 3. Our single match model trained with textual supervision, TSASPIRE consistently outperforms baselines. 4. Finally, our multi-match model OTASPIRE, while outperforming baselines sees aggregate performance similar to single match methods. This is reasonable given the aspect-specific annotation of

[3] MPNET-1B: https://bit.ly/2Zbm2Iq

[4] Initialization indicated via subscript in tables.

| CSFCUBE facets → | *Aggregated* | | *Background* | | *Method* | | *Result* | |
|---|---|---|---|---|---|---|---|---|
| Models | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ |
| MPNET-1B | 34.64 | 54.94 | 41.06 | 65.86 | 27.21 | 42.48 | 36.07 | 54.94 |
| SENTBERT-PP | 26.77 | 48.57 | 35.43 | 60.80 | 16.19 | 33.40 | 29.16 | 48.57 |
| SENTBERT-NLI | 25.23 | 45.39 | 30.78 | 54.23 | 15.02 | 31.10 | 30.27 | 45.39 |
| UNSIMCSE-BERT | 24.45 | 42.59 | 30.03 | 51.59 | 14.82 | 31.23 | 28.76 | 42.59 |
| SUSIMCSE-BERT | 23.24 | 43.45 | 30.52 | 55.22 | 13.99 | 30.88 | 25.58 | 43.45 |
| CoSentBert | 28.95 | 50.68 | 35.78 | 61.27 | 19.27 | 38.77 | 32.15 | 50.68 |
| ICTSENTBERT | 28.61 | 48.13 | 35.93 | 59.80 | 15.62 | 35.91 | 34.42 | 48.13 |
| OTMPNET-1B | 36.41 | 56.91 | 43.23 | 67.60 | 28.69 | 43.49 | 37.76 | 60.30 |
| SPECTER | 34.23 | 53.28 | 43.95 | 66.70 | 22.44 | 37.41 | 36.79 | 56.67 |
| SPECTER-COCITE$_{\text{Scib}}$ | 37.90 | 58.16 | 48.40 | 68.71 | 26.95 | 46.79 | 38.93 | 59.68 |
| SPECTER-COCITE$_{\text{Spec}}$ | 37.39 | 58.38 | 49.99 | 70.03 | 25.60 | 45.99 | 37.33 | 59.95 |
| TSASPIRE$_{\text{Spec}}$ | 40.26 | 60.71 | 49.58 | 70.22 | **28.86** | **48.20** | 42.92 | 64.39 |
| OTASPIRE$_{\text{Spec}}$ | **40.79** | **61.41** | 50.56 | **71.04** | 27.64 | 46.46 | **44.75** | **67.38** |
| TS+OTASPIRE$_{\text{Spec}}$ | 40.26 | 60.86 | **51.79** | 70.99 | 26.68 | 47.60 | 43.06 | 64.82 |

Table 1: Test set results for baseline and proposed methods on CSFCUBE, an expert annotated fine-grained similarity dataset of computer science papers. Our approaches outperform strong prior models OT/MPNET-1B and SPECTER, by 5-6 points aggregated across queries. Metrics (MAP, NDCG$_{\%20}$) are computed based on a 2-fold cross-validation and averaged over three re-runs of models. Here, TSASPIRE: Text supervised single-match method, OTASPIRE: Optimal Transport multi-match method and TS+OTASPIRE: Multi-task multi aspect method.

CSFCUBE where we expect gains from modeling single fine-grained (contextualized) matches rather than aggregating multiple matches.

Now, we examine facet-specific performance: 1. Performance on `background` sees higher performance in general and the smallest gains for the proposed approaches. This may be attributed to `background` similarity being captured in *coarse*-grained topical similarity, a trait largely captured in existing baselines. 2. `method` similarity in CSFCUBE presents significant challenges (Mysore et al., 2021, Sec 6) since it relies upon procedural similarities across steps of a method and on domain knowledge based similarities - this is often captured in co-citation data (Fig 1 presents one such complex paraphrase example). We see strongest performance for TSASPIRE here. 3. Finally, given that paper results interpretations are often dependent on all aspects of a given paper, `result` similarity often depends on similarity across the whole abstract. This leads OTASPIRE which models multiple matches to see strong performance.

**Whole-abstract similarity:** Table 2 presents results on TRECCOVID$_{\text{RF}}$ and RELISH. At the outset, we note that while being annotated for whole-abstract relevance, these datasets present different characteristics. While TRECCOVID$_{\text{RF}}$ presents queries centered on a very specific topic, RELISH presents a much more diverse set of queries. Further, TRECCOVID$_{\text{RF}}$ pairs queries with pools of about 9000 candidates while RELISH has about 60 candidates per query. Next, we examine baselines.

| Models | TRECCOVID$_{\text{RF}}$ | | RELISH | |
|---|---|---|---|---|
| | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ |
| MPNET-1B | 17.35 | 43.87 | 52.92 | 69.69 |
| SENTBERT-PP | 11.12 | 34.85 | 50.80 | 67.35 |
| SENTBERT-NLI | 13.43 | 40.78 | 47.02 | 63.56 |
| UNSIMCSE-BERT | 9.85 | 34.27 | 45.79 | 62.02 |
| SUSIMCSE-BERT | 11.50 | 37.17 | 47.29 | 63.93 |
| CoSentBert | 12.80 | 38.07 | 50.04 | 66.35 |
| ICTSENTBERT | 9.80 | 33.62 | 47.72 | 63.71 |
| OTMPNET-1B | 27.46 | 58.70 | 57.46 | 74.64 |
| SPECTER | 28.24 | 59.28 | 60.62 | 77.20 |
| SPECTER-COCITE$_{\text{Scib}}$ | 30.60 | 62.07 | 61.43 | 78.01 |
| SPECTER-COCITE$_{\text{Spec}}$ | 28.59 | 60.07 | 61.43 | 77.96 |
| TSASPIRE$_{\text{Spec}}$ | 26.24 | 56.55 | 61.29 | 77.89 |
| OTASPIRE$_{\text{Spec}}$ | **30.92** | **62.23** | 62.57 | 78.95 |
| TS+OTASPIRE$_{\text{Spec}}$ | 30.90 | 62.18 | **62.71** | **79.18** |

Table 2: Test set results for baseline and proposed methods on TRECCOVID$_{\text{RF}}$ and RELISH, expert annotated abstract similarity datasets of biomedical papers. Our approaches outperform a strong prior model, SPECTER, by 2-3 points across metrics (MAP, NDCG$_{\%20}$). These are computed as averages over three model re-runs. Method names map similarly to Table 1.

1. In contrast to fine-grained similarity datasets the best sentence level model MPNET-1B, significantly underperforms an abstract level model, SPECTER, indicating the need for whole abstract representations for these datasets. Aggregating sentence matches as in OTMPNET-1B, drastically improves MPNET-1B. 2. Next, similar to results in Table 1, a model identical to SPECTER, but trained on co-citations, SPECTER-COCITE$_{\text{Spec}}$, outper-

forms SPECTER indicating the value of co-citation signal for whole-abstract similarity too.

In examining performance of our proposed methods, we note the following: 1. Across datasets, our method for single matches, TSASPIRE, outperforms context-independent sentence baselines by several points indicating the value of contextualization. However, this method still underperforms abstract-level baselines. 2. However, methods modeling multiple matches, OTASPIRE and TS+OTASPIRE, substantially outperform TSASPIRE as well as baseline prior work SPECTER and OTMPNET-1B. This performance indicates the strength of OT based aggregation of fine-grained matches for abstract level similarity.

We present results demonstrating the value of the proposed approach on the SCIDOCS benchmark in Appendix G. Further, we also present a set of ablations in Appendix H. These ablations establish the value of textual supervision over the encoder ($\text{BERT}_{\mathcal{E}}$) used for encoding the text, the value of optimal transport compared to attention alternatives, and alternative single-match models trained without co-citation contexts.

## 5 Related Work

*Aspect-based paper representations:* A large body of work learns structured representations of scientific papers. Jain et al. (2018) present an approach which learns pre-defined aspect (PICO) encoders for biomedical papers. Similarly work of Neves et al. (2019), Chan et al. (2018), and Kobayashi et al. (2018) each label paper texts and then compute aspect-specific embeddings for document classification or ranking using existing methods. This line of work often relies on pre-defined aspects and building aspect-specific methods. Our work leverages co-citation contexts to supervise free-text aspects with a new model, that is also not tied to a specific schema of labels.

*Fine-grained document representations:* Another similar line of work is modeling fine-grained document-document similarity at the level of words or latent topics. Examples include early work El-Arini and Guestrin (2011) presenting paper recommendation methods with unigram-level similarity between papers using authorship and citation links or using latent document topics (Gong et al., 2018; Yurochkin et al., 2019; Dieng et al., 2020).

Our approach represents documents via sentences, a common and intuitive structure for reasoning about scientific document facets (Chan et al., 2018; Zhou et al., 2020). Ginzburg et al. (2021) present a self-supervised model for contextual sentence representations in long documents similar to our ICT baseline (Lee et al., 2019).

*Ad-hoc Search:* A range of recent work in information retrieval presents multi-vector models intended to capture different aspects of candidate documents with score aggregation relying on summations, max, or attention functions (Khattab and Zaharia, 2020; Luan et al., 2021; Humeau et al., 2020), these however focusing on short-text queries seen in search or question answering (QA). Mitra et al. (2017) explore an approach to model term-level fine-grained similarities with neural networks, Liu et al. (2018) model fine-grained matches at the level of entity spans, and Akkalyoncu Yilmaz et al. (2019) model document relevance by aggregating sentence relevance. Similarly, recent work of Lee et al. (2021) models fine-grained matches for QA at the phrase level. Importantly, these methods rely on supervision from knowledge bases or QA datasets, limiting applicability to specific span definitions and areas with these resources, often not present in the scientific domain (Hope et al., 2021a).

A range of modeling approaches in the context of other tasks resemble elements of our approach. We describe these in Appendix J.

## 6 Conclusions

We presented ASPIRE, a scientific document similarity model that is trained by leveraging co-citation contexts for learning fine-grained similarity. We use co-citation contexts as a novel form of *textual* supervision to guide the learning of multi-vector document representations. Our model outperformed strong baselines on seven document similarity tasks across four English scientific text datasets. Moreover, we showed that a fast single-match method achieves competitive results, enabling fine-grained document similarity in large-scale scientific corpora. A future direction is the interactive use of our methods, with a system allowing users to highlight specific aspects of papers and retrieve contextually-relevant matches. Another promising application is for finding analogies — structural matches between texts describing ideas, as in scientific papers, to boost discovery (Hope et al., 2017, 2021b; Chan et al., 2018).

## References


Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-domain modeling of sentence-level evidence for document retrieval. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3490–3496, Hong Kong, China. Association for Computational Linguistics.

Alexandr Andoni, Piotr Indyk, and Ilya Razenshteyn. 2018. Approximate nearest neighbor search in high dimensions. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pages 3287–3318.

Stefanos Angelidis and Mirella Lapata. 2018. Multiple instance learning networks for fine-grained sentiment analysis. *Transactions of the Association for Computational Linguistics*, 6:17–31.

Arturs Backurs, Yihe Dong, Piotr Indyk, Ilya Razenshteyn, and Tal Wagner. 2020. Scalable nearest neighbor search for optimal transport. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 497–506. PMLR.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation.

Mark Berger, Jakub Zavrel, and Paul Groth. 2020. Effective distributed representations for academic expert search. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 56–71, Online. Association for Computational Linguistics.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. *arXiv preprint arXiv:1802.08301*.

Peter Brown, RELISH Consortium, and Yaoqi Zhou. 2019. Large expert-curated database for benchmarking document similarity detection in biomedical literature search. *Database*, 2019. Baz085.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. Solvent: A mixed initiative system for finding analogies between research papers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–21.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Khalid El-Arini and Carlos Guestrin. 2011. Beyond keyword search: Discovering relevant scientific literature. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, page 439–447, New York, NY, USA. Association for Computing Machinery.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. *ACL*.

Dvir Ginzburg, Itzik Malkiel, Oren Barkan, Avi Caciularu, and Noam Koenigstein. 2021. Self-supervised document similarity ranking via contextualized language models and hierarchical inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3088–3098, Online. Association for Computational Linguistics.

Bela Gipp and Jöran Beel. 2009. Citation proximity analysis (cpa): A new approach for identifying related work based on co-citation analysis. In *ISSI'09: 12th international conference on scientometrics and informetrics*, pages 571–575.

Hongyu Gong, Tarek Sakakini, Suma Bhat, and JinJun Xiong. 2018. Document similarity for texts of varying lengths via hidden topics. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2341–2351, Melbourne, Australia. Association for Computational Linguistics.

Braden Hancock, Paroma Varma, Stephanie Wang, Martin Bringmann, Percy Liang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1884–1895, Melbourne, Australia. Association for Computational Linguistics.

Tom Hope, Aida Amini, David Wadden, Madeleine van Zuylen, Sravanthi Parasa, Eric Horvitz, Daniel Weld, Roy Schwartz, and Hannaneh Hajishirzi. 2021a. Extracting a knowledge base of mechanisms from COVID-19 papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4489–4503, Online. Association for Computational Linguistics.

Tom Hope, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2017. Accelerating innovation through analogy mining. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 235–243, New York, NY, USA. Association for Computing Machinery.

Tom Hope and Dafna Shahaf. 2016. Ballpark learning: Estimating labels from rough group comparisons. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 299–314. Springer.

Tom Hope and Dafna Shahaf. 2018. Ballpark crowdsourcing: The wisdom of rough group comparisons. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 234–242.

Tom Hope, Ronen Tamari, Hyeonsu Kang, Daniel Hershcovich, Joel Chan, Aniket Kittur, and Dafna Shahaf. 2021b. Scaling creative inspiration with fine-grained functional facets of product ideas.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Maximilian Ilse, Jakub Tomczak, and Max Welling. 2018. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR.

Sarthak Jain, Edward Banner, Jan-Willem van de Meent, Iain J Marshall, and Byron C Wallace. 2018. Learning disentangled representations of texts with application to biomedical abstracts. In *EMNLP*, volume 2018, page 4683.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. In *BMC bioinformatics*, volume 12, pages 1–10. BioMed Central.

Yuta Kobayashi, Masashi Shimbo, and Yuji Matsumoto. 2018. Citation recommendation using distributed representation of discourse facets in scientific articles. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, JCDL '18, page 243–251, New York, NY, USA. Association for Computing Machinery.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 957–966, Lille, France. PMLR.

Dan Lahav, Jon Saad Falcon, Bailey Kuehl, Sophie Johnson, Sravanthi Parasa, Noam Shomron, Duen Horng Chau, Diyi Yang, Eric Horvitz, Daniel S Weld, et al. 2021. A search engine for discovery of scientific challenges and directions. *arXiv preprint arXiv:2108.13751*.

Jinhyuk Lee, Alexander Wettig, and Danqi Chen. 2021. Phrase retrieval learns passage retrieval, too. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3661–3672, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2021. Pretrained transformers for text ranking: Bert and beyond. *Synthesis Lectures on Human Language Technologies*, 14(4):1–325.

Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2019. Example-based search: a new frontier for exploratory search. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1411–1412.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2018. Entity-duet neural ranking: Understanding the role of knowledge graph semantics in neural information retrieval. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2395–2405, Melbourne, Australia. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

David Mimno and Andrew McCallum. 2007. Expertise modeling for matching papers with reviewers.

In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 500–509.

Bhaskar Mitra, Fernando Diaz, and Nick Craswell. 2017. Learning to match using local and distributed representations of text for web search. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1291–1299, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Shikhar Murty, Pang Wei Koh, and Percy Liang. 2020. ExpBERT: Representation engineering with natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2106–2113, Online. Association for Computational Linguistics.

Sheshera Mysore, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. 2021. CSFCube - a test collection of computer science research articles for faceted query by example. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Mariana Neves, Daniel Butzke, and Barbara Grune. 2019. Evaluation of scientific elements for text similarity in biomedical publications. In *Proceedings of the 6th Workshop on Argument Mining*, Florence, Italy. Association for Computational Linguistics.

Allen Nie, Arturo L. Pineda, Matt W. Wright, Hannah Wand, Bryan Wulf, Helio A. Costa, Ronak Y. Patel, Carlos D. Bustamante, and James Zou. 2020. *LitGen: Genetic Literature Recommendation Guided by Human Explanations*, pages 67–78.

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

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.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *NeurIPS 2020*. ACM.

Yale Song and Mohammad Soleymani. 2019. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kyle Swanson, Lili Yu, and Tao Lei. 2020. Rationalizing text matching: Learning sparse alignments via optimal transport. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5609–5626, Online. Association for Computational Linguistics.

Derek Tam, Nicholas Monath, Ari Kobren, Aaron Traylor, Rajarshi Das, and Andrew McCallum. 2019. Optimal transport-based alignment of learned character representations for string similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5907–5917, Florence, Italy. Association for Computational Linguistics.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: Constructing a pandemic information retrieval test collection. *SIGIR Forum*.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The COVID-19 open research dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M Solomon. 2019. Hierarchical optimal transport for document representation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Hamed Zamani, Mostafa Dehghani, W. Bruce Croft, Erik Learned-Miller, and Jaap Kamps. 2018. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 497–506, New York, NY, USA. Association for Computing Machinery.

Xuhui Zhou, Nikolaos Pappas, and Noah A. Smith. 2020. Multilevel text alignment with cross-document attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5012–5025, Online. Association for Computational Linguistics.

## A Co-citation Data

As noted in §3.1, we train the proposed methods on English co-cited papers. We build a dataset of co-cited papers from the S2ORC corpus[5] (Lo et al., 2020). Since our evaluation datasets draw on text from different domains we build training sets with co-cited papers for each: biomedicine for RELISH and TRECCOVID$_{\text{RF}}$, computer science

[5]Released under a CC BY-NC 2.0. license.

for CSFCUBE, and a 60/40 mix of biomedicine and CS for SCIDOCS. Each dataset contains 1.3M training triples.

Next we describe construction of our co-citation data given 8.1 million English full text articles in the S2ORC corpus which have been parsed for citation mentions and linked to cited papers in the corpus using automatic tools (Lo et al., 2020):

- Domain definition: We define our biomedical articles to be those tagged either "Medicine" or "Biology" in S2ORC. "Computer Science" tagged papers are treated as CS papers.
- Co-citation contexts: To obtain co-cotation contexts - we first obtain sentence boundaries for co-citation contexts using the `en_core_sci_sm` pipeline included in `spacy`.[6].
- Filtering abstracts: In selecting abstracts for our dataset we retain those that have a minimum of 3 sentences, and a maximum of 20 sentences. Further, abstracts where all the sentences are too small (3 tokens) are excluded. Similarly, abstracts with sentences greater than 80 tokens are excluded.
- Selecting training co-citated abstract data $\{\mathcal{P}\}$: Given contexts with qualifying abstracts as described above, we only retain co-citation contexts with 2 or 3 co-cited papers. A manual examination revealed that larger co-cited sets tended to be more loosely related.
- Selecting co-citation sentence training data for $\texttt{BERT}_{\mathcal{E}}$: Note that this represents a sentence encoder trained by treating co-citation contexts referencing the same paper as paraphrases. Here, we select co-citation contexts containing 2 or more co-cited papers as paraphrase sets $\mathcal{E}$.

Abstract level training triples for the biomedical and computer science sets are built by treating all unique pairs of papers as positives. 1.3 million triples were used for each domain - these were sampled from larger sets at random.

## B Evaluation Dataset Details

Here we provide further detail on the evaluation datasets overviewed in §4.

RELISH: An annotated dataset of biomedical abstract queries labelled by experts (Brown et al., 2019). In a number of cases expert annotators are the authors of query papers. Per query candidate pools are of size 60, with 1638 queries in development and test sets each. Dataset is released under a Creative Commons Attribution 4.0 International License.

$\text{TRECCOVID}_{\text{RF}}$: While the original TRECCOVID dataset of Voorhees et al. (2021) is labelled for ad-hoc search by experts, we reformulate the dataset for abstract similarity, treating all documents relevant to one ad-hoc query as similar to each other. From each original query and its respective relevance-labeled documents, we sample an abstract from relevant documents (relevance of 2) and use that as our query document. We treat all other relevant documents as positive examples for the query. Documents relevant for *other* queries are treated as irrelevant for the sampled query. This results in about 9000 candidates per query abstract in $\text{TRECCOVID}_{\text{RF}}$. $\text{TRECCOVID}_{\text{RF}}$ consists of about 1200 queries in the development and test splits each. This dataset builds on the CORD-19 dataset (Wang et al., 2020) released under a Apache License 2.0, the license of TRECCOVID however isn't clear from the dataset release.

SCIDOCS: A benchmark suite of tasks intended for abstract-level scientific document representations (Cohan et al., 2020). We evaluate our methods on the tasks of predicting: citations, co-citations, co-views, and co-reads. Per query candidate pools are of size 30 about 1000 queries per task and development and test split. We exclude classification and recommendation sub-tasks relying on additional inference components. Dataset is released under a GNU General Public License v3.0 license.

CSFCUBE: The dataset consists of 50 queries labelled for relevance against about 120 candidates per query. Dataset is released under a Creative Commons Attribution-NonCommercial 4.0 International license.

## C Co-citation Context Encoder

Here we present details of alternative design choices for our co-citation context encoder. In the use of $\texttt{BERT}_{\mathcal{E}}$, we note in §3.2.2 that this encoder is kept frozen during the course of training $\texttt{BERT}_{\theta}$. Fine-tuning $\texttt{BERT}_{\mathcal{E}}$ via a straight-through estimator (Bengio et al., 2013) under-performed freezing it. Using other encoders for scientific text such as SPECTER as $\texttt{BERT}_{\mathcal{E}}$ under-performed `CoSentBert`. A recent strong model for sentence representation `MPNet-1B`[7] lead to similar

[6] `https://allenai.github.io/scispacy/`

[7] `MPNet-1B`: `https://bit.ly/2Zbm2Iq`

performance on abstract and aspect-conditional tasks as CoSentBert, indicating that a minimum requisite sentence encoder is all that is needed for $\text{BERT}_{\mathcal{E}}$.

## D Baselines

Here we provide further detail on the baselines overviewed in §4.

MPNET-1B & OTMPNET-1B: A sentence level baseline of a MPNet (Song et al., 2020) base model, fine-tuned on 1.17 billion similar text pairs in a contrastive learning setup.[8] This training data broadly represents web and scientific texts. Further we combine MPNET-1B with an OT based aggregation scheme similar to our multi-match model to yield, OTMPNET-1B a baseline using optimal transport with a performant sentence encoder.

SimCSE: A recent sentence representation model (Gao et al., 2021). We compare to two model variants: an unsupervised model UNSIMCSE-BERT, and a variant supervised with NLI data, SUSIMCSE-BERT.

Sentence-Bert: A sentence level transformer model fine-tuned on similar sentence pairs (Reimers and Gurevych, 2019). We compare performance to two variants, SENTBERT-PP and SENTBERT-NLI, fine-tuned on paraphrases and natural language inference (NLI) data respectively.

CoSentBert: The sentence-level model we describe in §3.2.2: A SCIBERT model fine-tuned on co-citation sentence contexts referencing the same set of co-cited papers.

ICTSENTBERT: A SCIBERT sentence model trained using the self-supervised inverse close task (Lee et al., 2019). Here we train abstract sentence representations to capture the semantics of their paragraph contexts.

SPECTER: A state of the art abstract level representation (Cohan et al., 2020). Here a SCIBERT model is fine-tuned to maximize similarity between representations of *cited* papers. We also train a variant of this model on *co-cited* papers: SPECTER-COCITE.

For the baselines described above specific model names from the Hugging Face[9] and Sentence Transformers[10] libraries are as follows:

[8]MPNet-1B: https://bit.ly/2Zbm2Iq

[9]https://huggingface.co/

[10]https://www.sbert.net/docs/pretrained_models.html

MPNet-1B: HF; sentence-transformers/all-mpnet-base-v2.

SimCSE: HF; princeton-nlp/sup-simcse-bert-base-uncased, princeton-nlp/unsup-simcse-bert-base-uncased.

Sentence-Bert: ST; Paraphrases: paraphrase-TinyBERT-L6-v2. NLI: nli-roberta-base-v2 from the Sentence-Transformers library.

## E Training

All our approaches are trained using the Adam optimizer with an initial linear warm-up for 2000 steps followed by a linear decay using gradient accumulation for a batch size of 30. The margin $m$ in the triplet loss is set to 1. We implement all methods using PyTorch, HuggingFace, and GeomLoss libraries. Training convergence is established based on the loss on a held out set of co-citation data ensuring that training does not rely on a labelled dataset for convergence checks.

All experiments were run with data parallelism over servers nodes with the following GPU configurations: 8×12GB NVIDIA GeForce GTX 1080 Ti GPUs, 4×24GB NVIDIA Tesla M40 GPUs, or 2×48GB NVIDIA Quadro RTX 8000 GPUs. Servers had 12-24 CPUs per node and 256-385GB RAM. The training time per experiment varied from 5-20 hours, and the experiments in this paper represent about 4746 GPU hours of training.

## F Model Hyper-Parameters

Here we report the best performing model hyper-parameters. This is done per training dataset. For computer science trained models evaluated on CSFCUBE:

- Specter-$\text{CoCite}_{\texttt{Scib}}$: LR 2e-5.
- Specter-$\text{CoCite}_{\texttt{Spec}}$: LR 2e-5.
- $\text{TSASPIRE}_{\texttt{Spec}}$: LR 2e-5.
- $\text{OTASPIRE}_{\texttt{Spec}}$: LR 2e-5. $\tau$ 0.5.
- $\text{TS+OTASPIRE}_{\texttt{Spec}}$: LR 1e-5. $\tau$ 0.5.

For biomedical trained models evaluated on TRECCOVID and RELISH:

- Specter-$\text{CoCite}_{\texttt{Scib}}$: LR 2e-5.
- Specter-$\text{CoCite}_{\texttt{Spec}}$: LR 2e-5.
- $\text{TSASPIRE}_{\texttt{Spec}}$: LR 2e-5.
- $\text{OTASPIRE}_{\texttt{Spec}}$: LR 2e-5. $\tau$ 5000.
- $\text{TS+OTASPIRE}_{\texttt{Spec}}$: LR 2e-5. $\tau$ 5000.

For biomedical+computer science trained models evaluated on TRECCOVID and RELISH:

| SciDocs tasks → | Citations | | Co-Citations | | Co-Reads | | Co-Views | |
|---|---|---|---|---|---|---|---|---|
| Models | MAP | NDCG | MAP | NDCG | MAP | NDCG | MAP | NDCG |
| MPNET-1B | 86.76 | 92.63 | 85.68 | 92.16 | 83.45 | 90.47 | 82.51 | 89.29 |
| SPECTER | **92.39** | **95.90** | 88.32 | 93.88 | 86.42 | 92.39 | 84.65 | 90.70 |
| SPECTER-COCITE$_{\texttt{Scib}}$ | 89.16 ±0.33 | 93.97 ±0.28 | 90.21 ±0.18 | 94.76 ±0.14 | 86.85 ±0.22 | 92.51 ±0.18 | 85.70 ±0.16 | 91.37 ±0.09 |
| SPECTER-COCITE$_{\texttt{Spec}}$ | 89.85 ±0.10 | 94.26 ±0.08 | 90.82 ±0.17 | 95.11 ±0.11 | 87.14 ±0.14 | 92.65 ±0.13 | 85.81 ±0.10 | 91.35 ±0.05 |
| TSASPIRE$_{\texttt{Spec}}$ | 90.99 ±0.26 | 95.04 ±0.17 | **90.92** ±0.06 | **95.26** ±0.05 | 87.51 ±0.07 | 92.97 ±0.06 | **85.87** ±0.20 | **91.46** ±0.14 |
| OTASPIRE$_{\texttt{Spec}}$ | 91.13 ±0.28 | 95.08 ±0.20 | 90.88 ±0.13 | 95.25 ±0.02 | 87.50 ±0.14 | 92.90 ±0.12 | 85.70 ±0.20 | 91.30 ±0.11 |
| TS+OTASPIRE$_{\texttt{Spec}}$ | 91.09 ±0.33 | 95.03 ±0.17 | 90.83 ±0.08 | 95.22 ±0.05 | **87.60** ±0.05 | **92.98** ±0.01 | 85.81 ±0.25 | 91.42 ±0.15 |

Table 3: Test set results for baseline and proposed methods on sub-tasks included in the SCIDOCS benchmark. Our approaches outperform a prior strong model, SPECTER, by 1-1.5 points on 3 of 4 sub-tasks. Metrics (MAP, NDCG) are computed based on averages over three re-runs of models. SPECTER uses model parameters as part of the Huggingface library. Here, TSASPIRE: Text supervised single-match method, OTASPIRE: Optimal Transport multi-match method and TS+OTASPIRE: Multi-task multi aspect method.

- Specter-CoCite$_{\texttt{Scib}}$: LR 2e-5.
- Specter-CoCite$_{\texttt{Spec}}$: LR 2e-5.
- TSASPIRE$_{\texttt{Spec}}$: LR 1e-5.
- OTASPIRE$_{\texttt{Spec}}$: LR 1e-5. $\tau$ 5000.
- TS+OTASPIRE$_{\texttt{Spec}}$: LR 1e-5. $\tau$ 5000.

We found it beneficial to use a low temperature $\tau$ in computing distributions $\mathbf{x}$ for OT computation for CSFCUBE - a fine-grained similarity dataset. On the other hand we found it beneficial to use a high temperature $\tau$ in computing distributions $\mathbf{x}$, causing it to be effectively uniform, for OT computation in whole-abstract datasets SCIDOCS, RELISH, and TRECCOVID$_{\texttt{RF}}$. This is reasonable given the nature of similarity captured in these datasets. Hyper-parameters of the underlying encoders were not changed from their default values – other hyper-parameters are common to methods and desribed in §4.

Finally, in computing OT transport plans, we optimize a entropy regularized objective: $\min_{\mathbf{P}\in\mathcal{S}}\langle\mathbf{D},\mathbf{P}\rangle - \frac{1}{\lambda}h(\mathbf{P})$. Our experiments use a fixed value of $\lambda = 20$.

**Hyper-parameter tuning:** We tune the hyper-parameters of all the ablated and proposed methods across the different datasets on development set performance. For CSFCUBE the *Aggregated* dev set performance was used for computer science training data models, TRECCOVID$_{\texttt{RF}}$ and RELISH dev sets were used for biomedical data models with ties between the two broken by the more challenging TRECCOVID$_{\texttt{RF}}$ performance, and computer science +biomedical data models were tuned on average task performance of SCIDOCS tasks. Given the expense of training models (about 20h for the proposed models) we first tune softmax temperatures then tuned learning rates. Large changes across learning rates weren't observed for the models. All learning rates are tuned over the range {1e-5, 2e-5, 3e-5}, OT sentence softmax temperatures $\tau$ are tuned over {0.5, 1, 5, 5000}, and softmax temperatures for ablation A3 was tuned over {0.5, 1, 5}.

## G SCIDOCS Benchmark Result

**SciDocs Benchmark:** Table 3 indicates performance on the abstract level document similarity benchmark SCIDOCS of Cohan et al. (2020). First we note that the strong performance of SPECTER indicates a smaller gap to be closed. Here, although our proposed methods see similar performance to each other they consistently outperform SPECTER on 3 of 4 tasks establishing state of the art performance. Given SPECTER's citation training signal and our co-citation signal, we see better performance on the `Citations` and `Co-Citation` tasks respectively. Finally, note that our co-citation trained approaches broadly see better performance (1-1.5 points) on extrinsic tasks of `Co-Reads` and `Co-Views` indicating the value of this signal.

## H Ablations

Here we ablate a range of model components in establishing factors which contribute performance. In ablations we only report performance on CSFCUBE, TRECCOVID$_{\texttt{RF}}$, and RELISH.

**A1. Does TSASPIRE gain from textual supervision over the encoder used to compute alignment?** TSASPIRE relies upon a sentence alignment encoder, $\texttt{BERT}_{\mathcal{E}}$ in §3.2.2, to compute alignments,

| | CSFCube *Agg.* | | TRECCOVID$_{RF}$ | | RELISH | |
|---|---|---|---|---|---|---|
| | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ |
| ABSASPIRE$_{\text{Spec}}$ | 37.03 $\pm$1.39 | 59.57 $\pm$0.76 | 25.42 $\pm$0.9 | 55.34 $\pm$0.55 | 58.78 $\pm$0.69 | 75.80 $\pm$0.57 |
| TSASPIRE$_{\text{Spec}}$ | 40.26 $\pm$0.93 | 60.71 $\pm$0.67 | 26.24 $\pm$0.45 | 56.55 $\pm$0.65 | 61.29 $\pm$0.51 | 77.89 $\pm$0.42 |

Table 4: Results for Ablation A1. Performance of TSASPIRE trained with textual supervision from co-citation contexts ablated for the effect of the text vs. influence of the text encoder (BERT$_{\mathcal{E}}$=`CoSentBert`; in §3.2.2) used to compute alignments to the co-citation contexts. Standard deviation across 3 model re-runs under mean performance.

| | CSFCube *Agg.* | | TRECCOVID$_{RF}$ | | RELISH | |
|---|---|---|---|---|---|---|
| | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ |
| ATTASPIRE$_{\text{Spec}}$ | 41.85 $\pm$1.52 | 61.67 $\pm$0.82 | 29.51 $\pm$0.78 | 60.96 $\pm$0.51 | 61.92 $\pm$0.52 | 78.54 $\pm$0.50 |
| OTASPIRE$_{\text{Spec}}$ | 40.79 $\pm$0.53 | 61.41 $\pm$0.52 | 30.92 $\pm$0.53 | 62.23 $\pm$0.67 | 62.57 $\pm$0.29 | 78.95 $\pm$0.26 |

Table 5: Results for Ablation A2. Performance for an alternative method, ATTASPIRE, for modeling multiple matches with an attention mechanism instead of optimal transport in the proposed method. Standard deviation across 3 model re-runs under mean performance.

$\hat{i}_p, \hat{i}_{p'}$, from the co-citation context to the co-cited abstracts. Here we investigate if improvements in TSASPIRE are attributable to BERT$_{\mathcal{E}}$ or to the co-citation contexts themselves. We investigate this by comparing the performance of TSASPIRE to a model trained to maximize the alignment between abstract sentences directly computed using BERT$_{\mathcal{E}}$, we refer to this as ABSASPIRE. This may be viewed as a form of knowledge distillation where alignments from a more local sentence encoder model, BERT$_{\mathcal{E}}$, are distilled into the contextual sentence encoder of TSASPIRE. As Table 4 shows, TSASPIRE consistently outperforms ABSASPIRE, indicating the value added by natural language supervision from the co-citation contexts.

**A2. Can multi-aspect matching use attention aggregation instead of optimal transport?** Since our multi-aspect match model uses a soft sparse matching with optimal transport we examine contributions of this component by comparing performance of a model (ATTASPIRE) trained with soft-alignment using an attention mask, $\mathbf{A}$ – attention is also a popular choice in prior work Humeau et al. (2020); Zhou et al. (2020). Here, $f_{\texttt{Att}}(p, p') = \langle \mathbf{D}, \mathbf{A} \rangle$ with, $\mathbf{A} = \texttt{softmax}(-\mathbf{D}/\tau)$. Note that OT imposes specific inductive bias via the structure of the trasport plan in ensuring it to be a permutation matrix - a desirable property in computing multiple alignments between a set of points. Table 5 examines performance of these model variants. Broadly, ATTASPIRE sees performance comparable or worse than OTASPIRE. While ATTASPIRE sees improved performance in CSFCube it sees much larger variation across runs. In our abstract retrieval datasets, where we expect gains from modeling multiple matches, we see better or similar performance from OTASPIRE over ATTASPIRE.

| | CSFCube *Aggregated* | | TRECCOVID$_{RF}$ | | RELISH | |
|---|---|---|---|---|---|---|
| | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ |
| MAXASPIRE$_{\text{SciB}}$ | 36.66 $\pm$1.37 | 57.68 $\pm$0.86 | 24.87 $\pm$1.15 | 54.33 $\pm$1.49 | 61.36 $\pm$0.31 | 78.10 $\pm$0.24 |
| MAXASPIRE$_{\text{Spec}}$ | 39.42 $\pm$1.38 | 60.63 $\pm$1.53 | 25.84 $\pm$0.85 | 56.52 $\pm$1.21 | 61.20 $\pm$0.97 | 78.00 $\pm$0.36 |
| TSASPIRE$_{\text{SciB}}$ | 40.10 $\pm$0.76 | 60.92 $\pm$0.61 | 27.68 $\pm$0.71 | 58.42 $\pm$0.75 | 61.45 $\pm$0.31 | 78.12 $\pm$0.33 |
| TSASPIRE$_{\text{Spec}}$ | 40.26 $\pm$0.93 | 60.71 $\pm$0.67 | 26.24 $\pm$0.45 | 56.55 $\pm$0.65 | 61.29 $\pm$0.51 | 77.89 $\pm$0.42 |

Table 6: Results for Ablation A3. Performance of a simpler single-match model, MAXASPIRE, trained using only BERT$_\theta$ representations while also varying encoder initialization between SPECTER and SCIBERT (indicated as subscripts for models). Standard deviation across 3 model re-runs under mean performance.

**A3. Can single-match models be learned without co-citation contexts?** While our model for single matches leverages weak textual supervision from co-citation contexts, we ask if these models can be learned in the absence of this supervision. We answer this by training a simpler model, MAXASPIRE, which finds the maximally aligned aspects between documents using the representations from BERT$_\theta$ alone, giving us $f_{\texttt{Max}}(p, p') = \texttt{max}_{i,j}\mathbf{D}$. To examine the role of

BERT$_\theta$ we compare performance with different initializations, with SPECTER presenting a initial model fine-tuned for similarity vs SCIBERT which isnt fine-tuned for text similarity.

We note the following from the results in Table 6: MAXASPIRE sees a dependence on the underlying encoder, a SCIBERT initialization nearly always sees poorer performance – only seeing performance competitive with TSASPIRE when initialized with SPECTER. This is reasonable given that this model must bootstrap fine-grained similarity while only relying on the encoder induced similarity. In cases where MAXASPIRE matches performance of TSASPIRE it sees larger performance differences across runs which may also be explained by the dependence on the initialization. Finally, TSASPIRE consistently sees similar or better performance with varying initialization, indicating the value of our text supervised method.

## I Extended Results

Tables 1, 2 in §4.1 omit presentation of standard deviations across runs for the proposed approaches for brevity. We include these in Tables 7 and 8.

## J Extended Related Work

A range of modeling approaches in multi-instance learning, models leveraging textual supervision, and optimal transport resemble elements of our approach. We describe these next.

*Multi-instance Learning:* Our work applies MIL for learning fine-grained similarity, while prior work has most often been applied to classification or regression tasks (Hope and Shahaf, 2016, 2018; Ilse et al., 2018; Angelidis and Lapata, 2018). Our work bears resemblance to an application of MIL in content based image retrieval (Song and Soleymani, 2019), where MIL is applied to learn alignments between image and text aspects.

*Textual Supervision:* Our use of co-citation text as a source of textual supervision draws on other work leveraging textual justifications of labels as a source of supervision for classification tasks (Hancock et al., 2018; Murty et al., 2020) - co-citation contexts may be considered justifications for similarity of co-cited papers. Nie et al. (2020) presents work in a biomedical literature recommendation task, where human justifications of a relevance label are used to identify unigram features indicative of the label and train a recommendation model.

*Optimal Transport:* Our use of optimal transport draws on other recent work in learning sparse alignments between texts (Swanson et al., 2020; Tam et al., 2019). Work of Swanson et al. (2020) learns sparse *binary* alignments for sentence and document similarity tasks to rationalize decisions and Tam et al. (2019) leverage sparse soft alignments between characters for string similarity. Kusner et al. (2015) uses alignment based on word embeddings for document classification tasks using a K-nearest neighbors method. However, applying OT at the word level in scientific documents would lead to a large increase in computational complexity.

| CSFCube facets→ | *Aggregated* | | background | | method | | result | |
|---|---|---|---|---|---|---|---|---|
| | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ |
| MPNet-1B | 34.64 | 54.94 | 41.06 | 65.86 | 27.21 | 42.48 | 36.07 | 54.94 |
| SentBert-PP | 26.77 | 48.57 | 35.43 | 60.80 | 16.19 | 33.40 | 29.16 | 48.57 |
| SentBert-NLI | 25.23 | 45.39 | 30.78 | 54.23 | 15.02 | 31.10 | 30.27 | 45.39 |
| UnSimCSE-BERT | 24.45 | 42.59 | 30.03 | 51.59 | 14.82 | 31.23 | 28.76 | 42.59 |
| SuSimCSE-BERT | 23.24 | 43.45 | 30.52 | 55.22 | 13.99 | 30.88 | 25.58 | 43.45 |
| CoSentBert | 28.95 | 50.68 | 35.78 | 61.27 | 19.27 | 38.77 | 32.15 | 50.68 |
| ICTSentBert | 28.61 | 48.13 | 35.93 | 59.80 | 15.62 | 35.91 | 34.42 | 48.13 |
| otMPNet-1B | 36.41 | 56.91 | 43.23 | 67.60 | 28.69 | 43.49 | 37.76 | 60.30 |
| Specter | 34.23 | 53.28 | 43.95 | 66.70 | 22.44 | 37.41 | 36.79 | 56.67 |
| Specter-CoCite$_{Scib}$ | 37.90 ±1.48 | 58.16 ±1.9 | 48.40 ±2.51 | 68.71 ±2.71 | 26.95 ±0.96 | 46.79 ±0.74 | 38.93 ±2.17 | 59.68 ±3.58 |
| Specter-CoCite$_{Spec}$ | 37.39 ±0.73 | 58.38 ±0.86 | 49.99 ±1.2 | 70.03 ±1.16 | 25.60 ±0.53 | 45.99 ±1.35 | 37.33 ±0.86 | 59.95 ±1.02 |
| tsAspire$_{Spec}$ | 40.26 ±0.93 | 60.71 ±0.67 | 49.58 ±1.59 | 70.22 ±1.74 | 28.86 ±1.71 | 48.20 ±1.72 | 42.92 ±0.54 | 64.39 ±0.28 |
| otAspire$_{Spec}$ | **40.79** ±0.53 | **61.41** ±0.52 | 50.56 ±1.52 | 71.04 ±1.42 | 27.64 ±0.92 | 46.46 ±0.1 | **44.75** ±1.57 | **67.38** ±0.99 |
| ts+otAspire$_{Spec}$ | 40.26 ±0.71 | 60.86 ±0.58 | **51.79** ±1.18 | 70.99 ±1.28 | 26.68 ±3.21 | 47.60 ±2.45 | 43.06 ±0.21 | 64.82 ±0.19 |

Table 7: Test set results for baseline and proposed methods on CSFCube, an expert annotated fine-grained similarity dataset of computer science papers. Our approaches outperform strong prior models ot/MPNet-1B and Specter, by 5-6 points aggregated across queries. Metrics (MAP, NDCG$_{\%20}$) are computed based on a 2-fold cross-validation and averaged over three re-runs of models. Standard deviations are below run averages. Here, tsAspire: Text supervised single-match method, otAspire: Optimal Transport multi-match method and ts+otAspire: Multi-task multi aspect method.

| | TRECCOVID$_{RF}$ | | RELISH | |
|---|---|---|---|---|
| | MAP | NDCG$_{\%20}$ | MAP | NDCG$_{\%20}$ |
| MPNet-1B | 17.35 | 43.87 | 52.92 | 69.69 |
| SentBert-PP | 11.12 | 34.85 | 50.80 | 67.35 |
| SentBert-NLI | 13.43 | 40.78 | 47.02 | 63.56 |
| UnSimCSE-BERT | 9.85 | 34.27 | 45.79 | 62.02 |
| SuSimCSE-BERT | 11.50 | 37.17 | 47.29 | 63.93 |
| CoSentBert | 12.80 | 38.07 | 50.04 | 66.35 |
| ICTSentBert | 9.80 | 33.62 | 47.72 | 63.71 |
| otMPNet-1B | 27.46 | 58.70 | 57.46 | 74.64 |
| Specter | 28.24 | 59.28 | 60.62 | 77.20 |
| Specter-CoCite$_{Scib}$ | 30.60 ±0.87 | 62.07 ±0.95 | 61.43 ±0.32 | 78.01 ±0.1 |
| Specter-CoCite$_{Spec}$ | 28.59 ±0.25 | 60.07 ±0.36 | 61.43 ±0.24 | 77.96 ±0.23 |
| tsAspire$_{Spec}$ | 26.24 ±0.45 | 56.55 ±0.65 | 61.29 ±0.51 | 77.89 ±0.42 |
| otAspire$_{Spec}$ | **30.92** ±0.53 | **62.23** ±0.67 | 62.57 ±0.29 | 78.95 ±0.26 |
| ts+otAspire$_{Spec}$ | 30.90 ±0.71 | 62.18 ±0.7 | **62.71** ±0.16 | **79.18** ±0.15 |

Table 8: Test set results for baseline and proposed methods on TRECCOVID$_{RF}$ and RELISH, expert annotated abstract similarity datasets of biomedical papers. Our approaches outperform a strong prior model, Specter, by 2-3 points across metrics (MAP, NDCG$_{\%20}$). These are computed as averages over three model re-runs. Standard deviations are below run averages. Method names map similarly to Table 7.