SPECTER2: Adapting scientific document embeddings to multiple fields and task formats

November 27, 2023

Amanpreet Singh | **Ai2**

Share

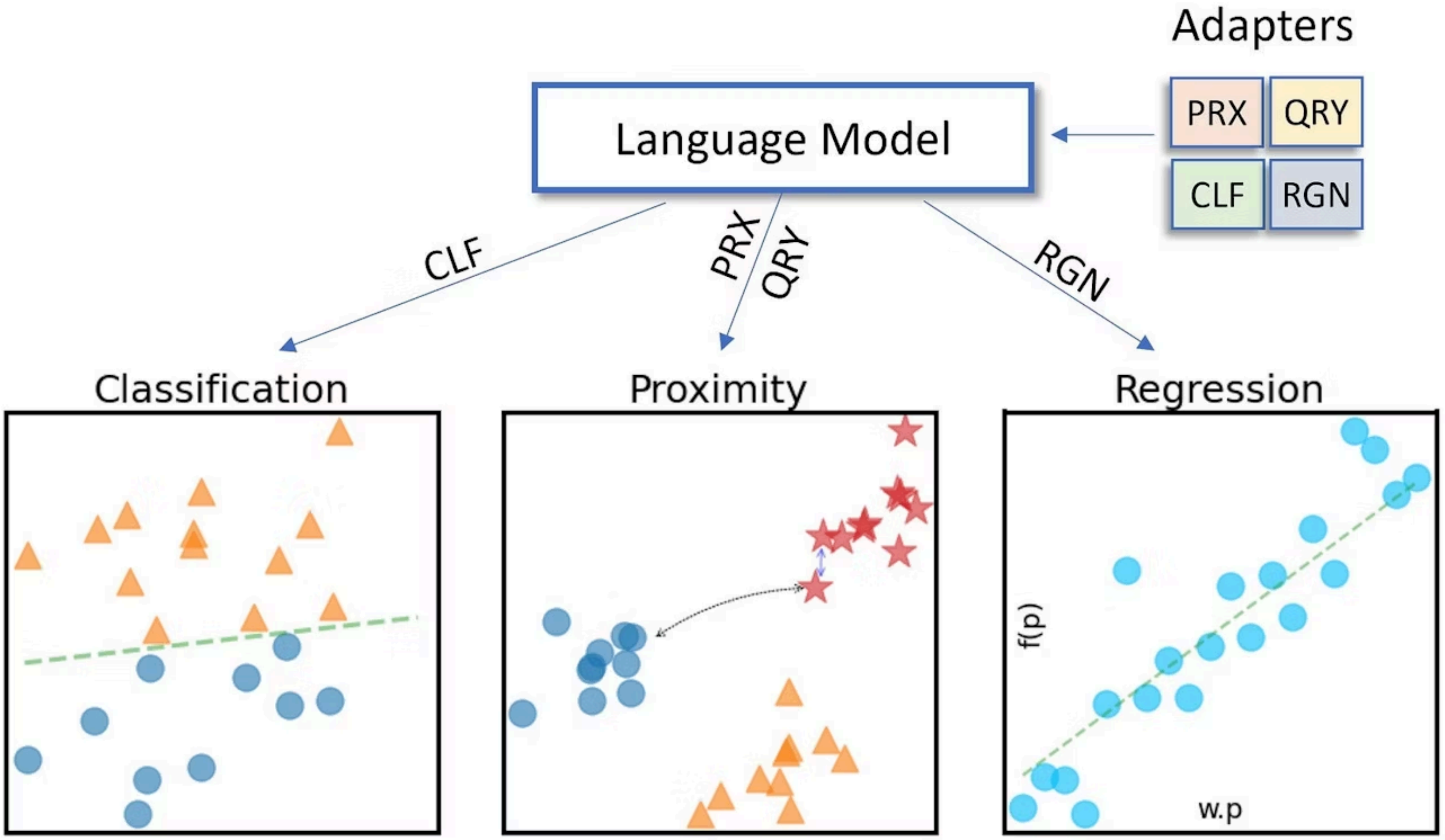


Fig 1: SPECTER2 uses adapters to generate task-specific embeddings for an input document

TL;DR: We create SPECTER2 ↗, a new scientific document embedding model via a 2-step training process on large datasets spanning 9 different tasks and 23 fields of study. The model improves upon its **predecessor** ↗ and can generate different embeddings based on the task type. We also release **SciRepEval** ↗, a comprehensive evaluation benchmark for scientific embeddings consisting of 24 tasks.

*For a more detailed description of our methodology and results, please refer to our paper-**SciRepEval: A Multi-Format Benchmark for Scientific Document Representations**↗ which has been accepted to appear in EMNLP 2023.*

Scientific Document Embeddings: A Background

Embedding text documents into vector representations can be useful for a variety of downstream tasks like recommendation, ranking, and clustering. These vectors are then used either to find similar documents or as features in a computationally cheap model. **Transformer models** ↗ like **BERT** ↗, which are pre-trained on large quantities of text, are the go-to approach these days for embedding text in a semantic space. A variety of such **embedding models are available** ↗ for users to choose from. However, our focus in this post is scientific documents, and, as we show later in this post, general-purpose embedding models may not be the best choice for specialized fields like Science.

Models like **SPECTER** ↗ and **SciNCL** ↗ are adept at embedding scientific documents as they are specifically trained so that papers close to each other in the citation network are close in the embedding space as well. For each of these models, the input paper text is represented by a combination of its title and abstract. SPECTER, released in 2020, supplies embeddings for a variety of our offerings at **Semantic Scholar** ↗-user research feeds, author name disambiguation, **paper clustering** ↗, and many more! Along with SPECTER, we also released **SciDocs** ↗-a benchmark of 7 tasks for evaluating the efficacy of scientific document embeddings. SciNCL, which came out last year, improved upon SPECTER by relying on nearest-neighbor sampling rather than hard citation links to generate training examples.

Opportunities for Improvement

Despite their impressive performance on SciDocs, and subsequent adoption by the research community, both SPECTER and SciNCL suffer from a few limitations:

1. Almost 70% of the training data for the models comes from Computer Science and BioMed, leading to poor performance in other fields of study as shown in the **MDCR benchmark** ↗ where BM-25 outperformed both the models.
2. Both the models are trained only on citation prediction, which is helpful for finding similar papers but might not yield the optimum results for tasks like classification, regression, and ad-hoc search.
3. Out of the 7 tasks, 4 are designed to evaluate document similarity. Also, we found the model performances on the SciDocs tasks in our experiments to be highly correlated (Figure 2). Thus, the evaluation is not diverse enough for a comprehensive evaluation of scientific document embeddings.

The above considerations provided us with the opportunity to research and come up with an improved model and evaluation benchmark. We name them SPECTER2 and SciRepEval respectively, and discuss more about each in the following sections.

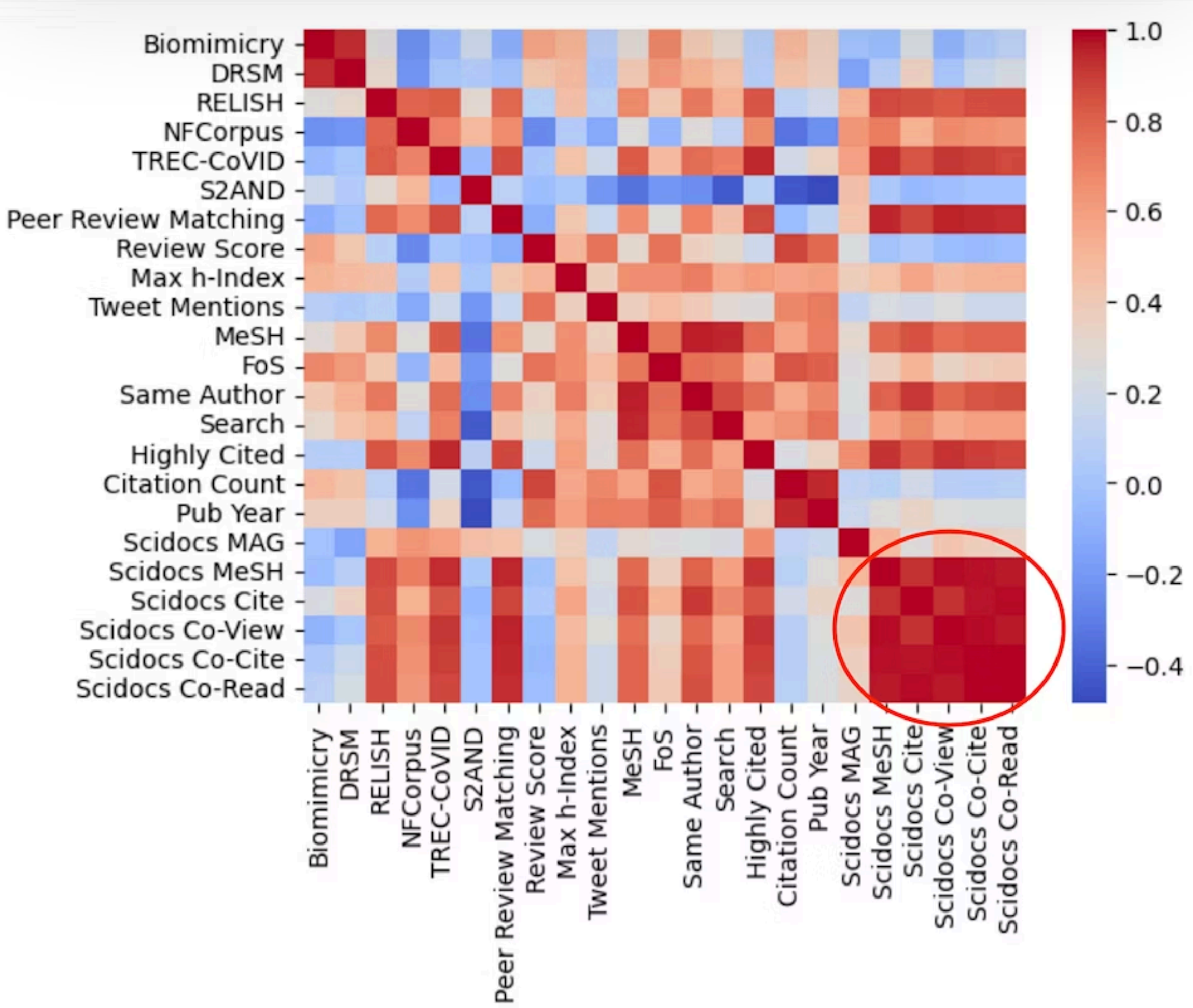


Fig 2: Pearson correlation plot between 18 different model evaluations across each pair of tasks we consider. SciDocs tasks (highlighted in the red circle) are all highly correlated with one another.

Modeling Deep Dive

SPECTER2 is trained in 2 steps, resulting in the ***SPECTER2 Base*** and ***SPECTER2 Adapters*** models respectively. Let's look at each of the steps in more detail:

Step 1: Pre-Training on Citation Prediction

Similar to its predecessor ↗, SPECTER2 is first pre-trained on the citation prediction task with a contrastive learning objective. We initialize with a **SciBERT** ↗ (BERT model trained on a large scientific corpus from scratch) checkpoint, and train it on triplet instances consisting of a query paper Q, with a positive P+ and negative P- candidate paper, such that the Euclidean distance d is lesser between Q and P+ as compared to Q and P-. Here, P+ is either cited by or cites Q; and, P- is either randomly sampled (easy -ve) or is cited by a citation of Q, but not by Q itself (hard -ve).

$$L_{triplet} = \max\{d(\mathcal{Q}_E, \mathcal{P}_E^+) - d(\mathcal{Q}_E, \mathcal{P}_E^-) + \epsilon, 0\}$$

This approach is very similar to how SPECTER was originally trained. The main difference here lies in the size and diversity of the training data. The new model, which we term SPECTER2 Base, is trained on 6M triplets, 10X more than what was used for SPECTER. Further, these triplets span 23 fields of study including not just Computer Science and Medicine, but also Art, Physics, Geography, and Geology, ensuring that we train on a diverse dataset.

Step 2: The Multiple Embeddings per Document Hypothesis

The previous step yields a good base model, but it is still trained on a single task only. **Multi-task training** ↗ is known to help models generalize well across tasks. So, we create 8 large-scale scientific document training tasks as part of **SciRepEval** ↗ belonging to four formats-classification, regression, proximity (similarity), and ad-hoc search as shown in Figure 1. However, simple multi-task led to worse results than the base model in our experiments as shown below in Table 1.

We hypothesized that a single embedding per document resulting from simple multi-task training without any task-specific parameters might not be expressive enough to generalize across multiple tasks. So, we introduce adapters specific to each task format for multi-task training. **Adapters** ↗ are linear modules attached to each layer of the transformer network, whose weights are adjusted during training while keeping the transformer fixed. Specific loss functions are optimized during training on the four task formats-*Cross Entropy* for classification, *Mean Squared Error* for regression, and the *Triplet Margin loss* for proximity/search. With the introduction of adapters, we can now obtain multiple different embeddings for the same document, tailored to the downstream task format. The model thus obtained is termed SPECTER2 Adapters.

Evaluation

As is expected, the resulting model is evaluated on held-out test sets from the training tasks; but we also wished to evaluate the generalizability of the embeddings. SciRepEval comes with an additional 16 out-of-training datasets for this purpose. We evaluate on not just the scientific document embedding models, but also a handful of recent general-purpose embedding models including OpenAI's *text-embedding-ada-002*.

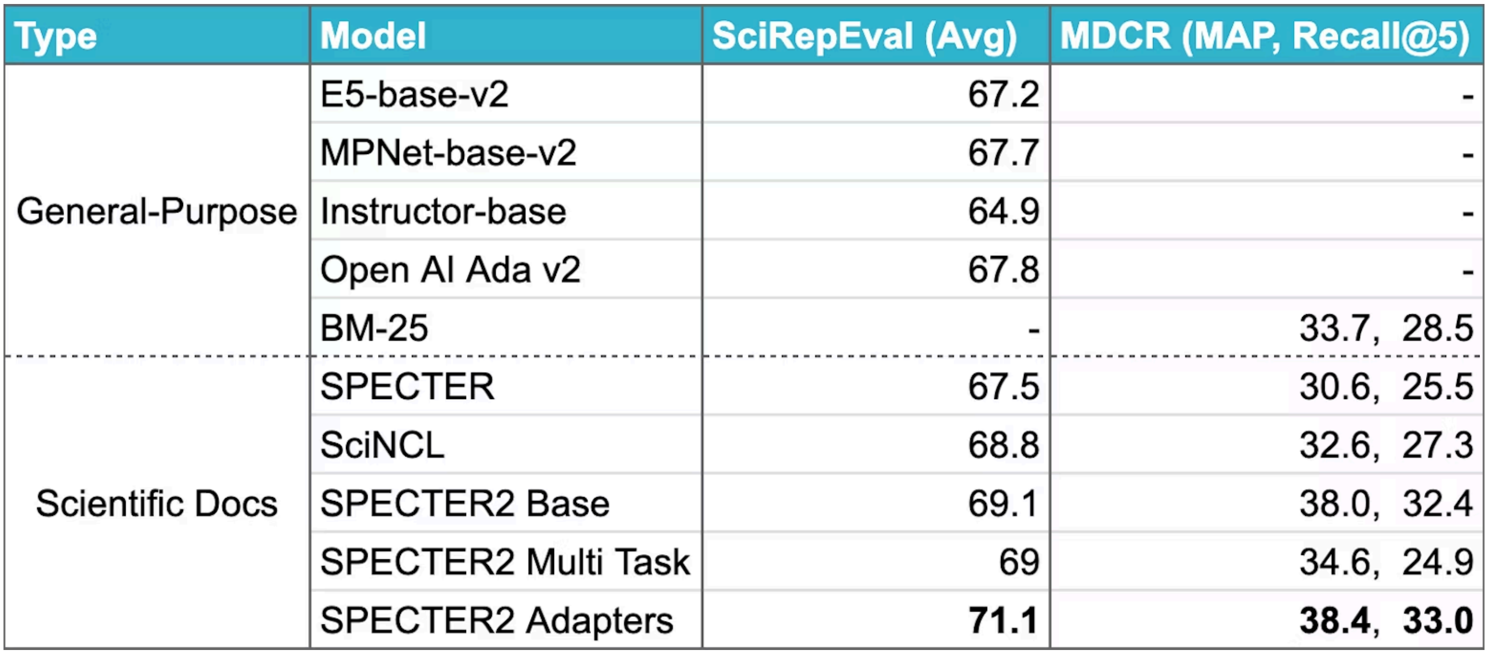
All the models in our experiments are evaluated on the 24 tasks from SciRepEval and the results are summarized in Table 1. Of those 24 tasks, 8 are new and we also include SciDocs as a subset.

The new tasks are briefly described below:

- MeSH ↗ Descriptors: Multi-class classification for 30 most frequently occurring top-level Medical Subject Heading (MeSH) descriptors, which are used for indexing biomedical research on PubMed ↗.
- FoS: Multi-label classification across 23 fields of study labels
- Citation count: Regression task for predicting log of citations of papers published over 5 years ago
- Year of Publication: Regression task for predicting year of publication for papers published after 2005

- h-Index of Authors: Regression task for predicting max h-Index among authors of a paper
- Tweet Mentions: Regression task to predict scaled number of mentions for traction of arXiv papers between 2010–2019.
- Highly Influential Citations: Re-ranking task for candidate papers, where a positive candidate paper B is cited more than 4 times by the query paper A
- Search: Retrieval task where candidate papers are returned for a short textual query

Overcoming the Limitations



Type	Model	SciRepEval (Avg)	MDCR (MAP, Recall@5)
General-Purpose	E5-base-v2	67.2	-
	MPNet-base-v2	67.7	-
	Instructor-base	64.9	-
	Open AI Ada v2	67.8	-
	BM-25	-	33.7, 28.5
Scientific Docs	SPECTER	67.5	30.6, 25.5
	SciNCL	68.8	32.6, 27.3
	SPECTER2 Base	69.1	38.0, 32.4
	SPECTER2 Multi Task	69	34.6, 24.9
	SPECTER2 Adapters	**71.1**	**38.4, 33.0**

Table 1: Evaluation results of models on SciRepEval and MDCR* benchmarks

A few observations from our summary results in Table 1 above:

1. The general purpose models - E5 ↗, Instructor ↗, and MPNet ↗ which are trained on large-scale Information Retrieval fall short on SciRepEval, which covers regression, classification, and re-ranking tasks along with retrieval.
2. Open AI embeddings ↗ are better than E5, Instructor, and MPNet but still worse than SPECTER2, and it costs about $200 to evaluate

them on SciRepEval.

3. Coming to scientific embedding models, simple multi-task training comes out to be worse than the base model on both SciRepEval and MDCR.
4. SPECTER 2 Adapters are better on average across all the SciRepEval tasks.

*Please note:

- About 23% of the papers (but only 0.03% citation links) from MDCR are also part of the SPECTER2 training data which might lead to some transductive advantage over the other scientific models.
- MPNet and E5 are trained on citation links from **S2ORC** ↗ -a corpus of 91M scientific papers. We found that this leads to a leakage of over 98% MDCR labels. Consequently, MDCR cannot be used to evaluate these models. A similar leakage analysis was not possible with Instructor and Open AI Ada embeddings without their source training data.

Public Release

Both SPECTER2 models and SciRepEval benchmark have been publicly released for users to build on. Please follow the below links for reproducibility and more info:

(i) Models on HuggingFace: **SPECTER Base** ↗, SPECTER Adapters- **classification** ↗, **regression** ↗, **proximity** ↗, **ad-hoc search** ↗ (**Note:** For general embedding tasks that don't fall under one of our formats, please use the **proximity** adapter termed SPECTER2 with the base model for best results.)

(ii) SPECTER2 **GitHub** ↗

(iii) The static paper embeddings are also available for free with the **Semantic Scholar public api** ↗ via the paper data end points with *embedding.specterv2* as the field parameter.

(iv) SciRepEval benchmark on HuggingFace: **https://huggingface.co/datasets/allenai/scirepeval** ↗ and **https://huggingface.co/datasets/allenai/scirepeval_test** ↗

(v) SciRepEval **GitHub** ↗

Acknowledgments

We would like to acknowledge our peers and mentors whose contributions were critical to SPECTER2 and SciRepEval:

- **Doug Downey** ↗
- **Sergey Feldman** ↗
- **Arman Cohan** ↗
- **Mike D'Arcy** ↗
- **Jonathan Bragg** ↗
- David Graham
- Zhipeng Hou

We would also like to acknowledge the support of NASA AATT Project for funding the **PeTaL** ↗ research and contributing the biomimicry dataset. This work was supported in part by NSF Grant 2033558.

First Name

Last Name

Email

Sign up →

Contact us

Questions about our work, or need support with one of our technologies?

Get in touch →

Resources

Media center
Documentation
Careers
Team directory

Community

Discord
Reddit

X/Twitter

GitHub

Hugging Face

LinkedIn

Bluesky

Threads

Legal

Terms of use

Privacy policy

DMCA policy

Business code of conduct

Responsible use