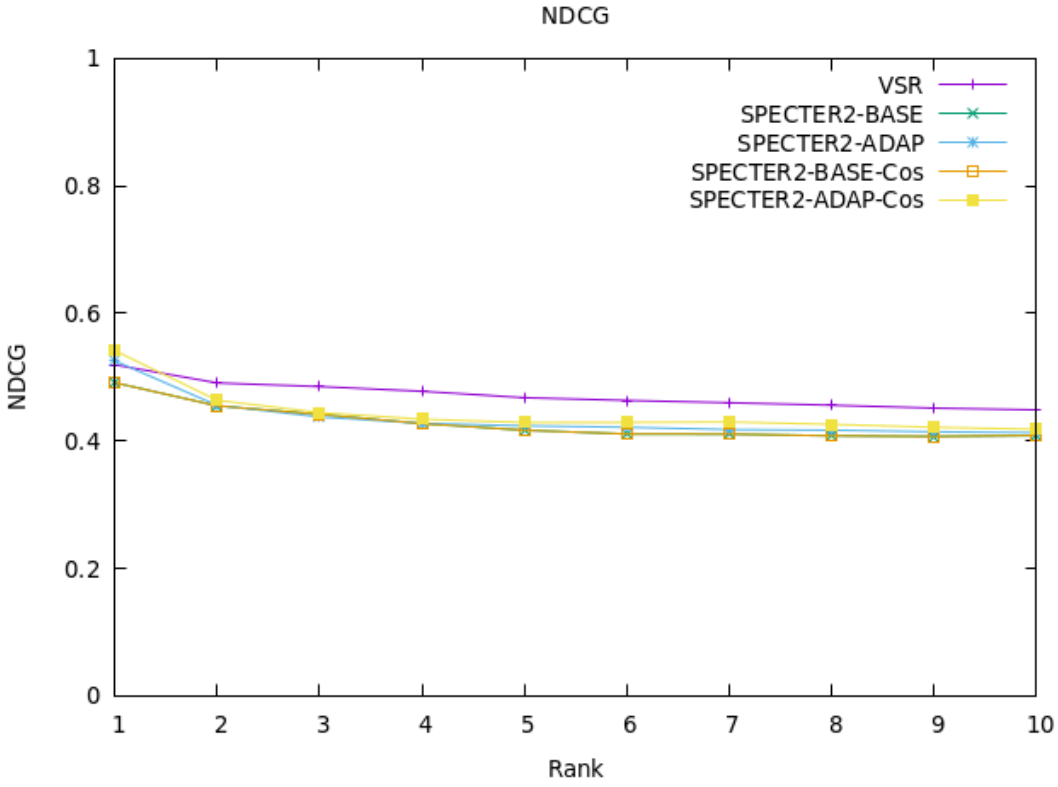
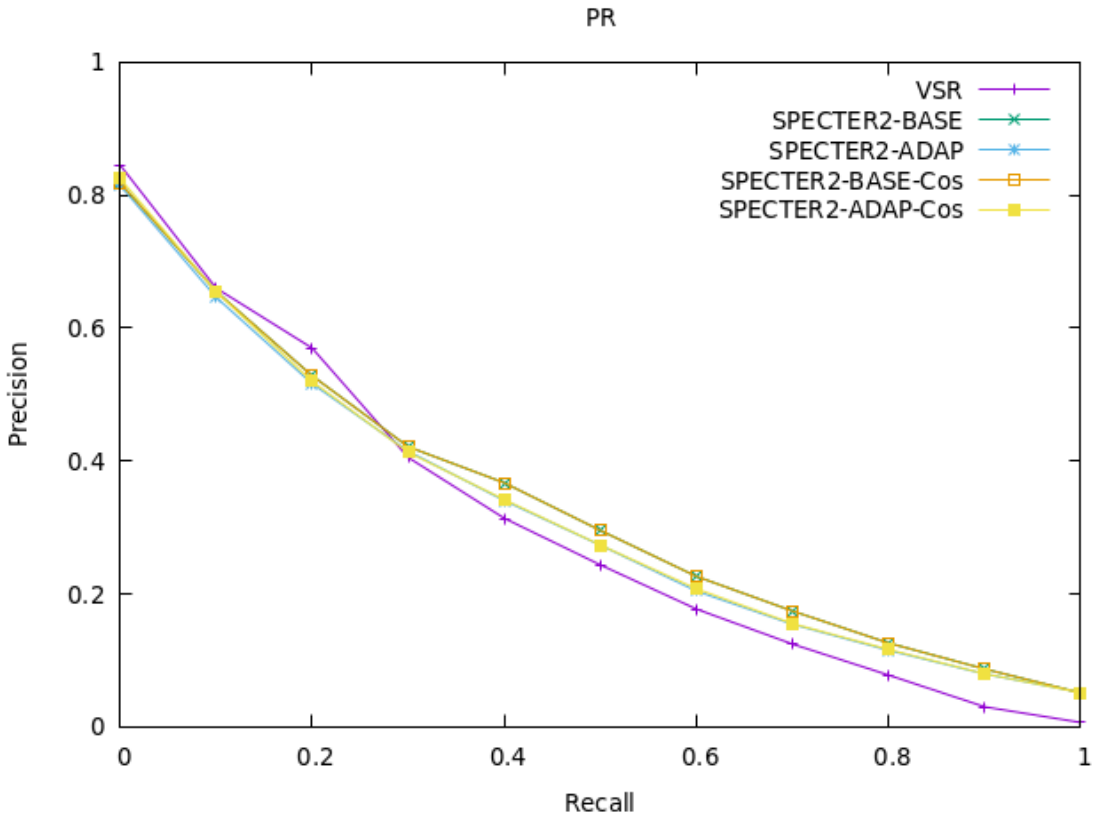
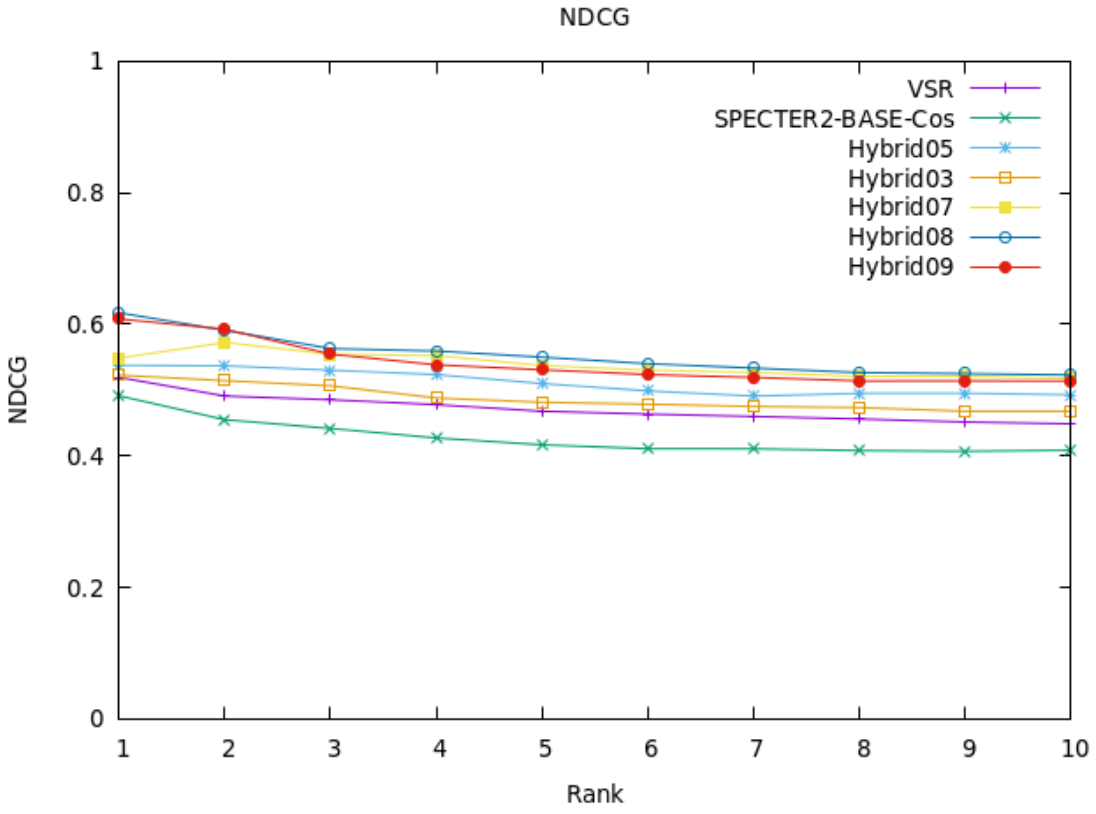
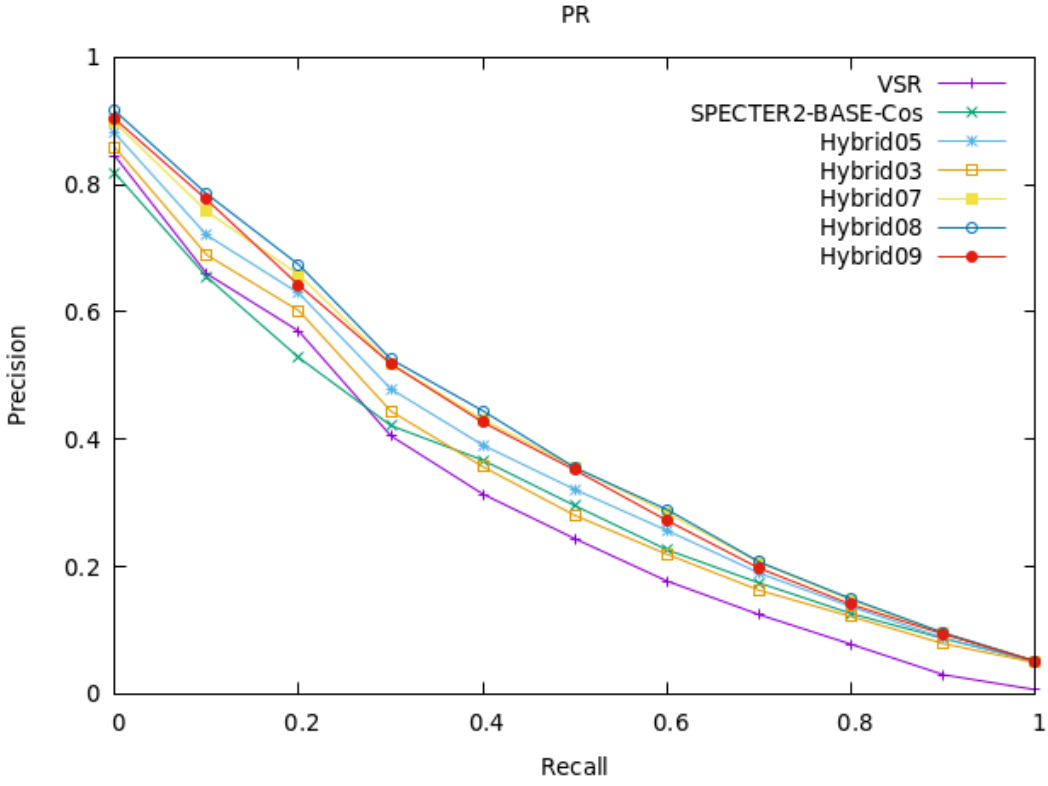

NDCG
VSR
SPECTER2-BASE
SPECTER2-ADAP
SPECTER2-BASE-Cos
SPECTER2-ADAP-Cos
NDCG
Rank
0
0.2
0.4
0.6
0.8
1
1
2
3
4
5
6
7
8
9
10


PR
VSR
SPECTER2-BASE
SPECTER2-ADAP
SPECTER2-BASE-Cos
SPECTER2-ADAP-Cos
Precision
Recall
0
0.2
0.4
0.6
0.8
1


NDCG
VSR
SPECTER2-BASE-Cos
Hybrid05
Hybrid03
Hybrid07
Hybrid08
Hybrid09
NDCG
Rank
0
0.2
0.4
0.6
0.8
1
1
2
3
4
5
6
7
8
9
10


PR
VSR
SPECTER2-BASE-Cos
Hybrid05
Hybrid03
Hybrid07
Hybrid08
Hybrid09
Precision
Recall
0
0.2
0.4
0.6
0.8
1

1. How does using the SPECTER2 embeddings compare to the VSR baseline in terms of retrieval accuracy (both PR and NDCG) ?

At lower levels of recall, SPECTER2 embeddings have lower precision values than VSR. However, at around 0.275 recall and higher, SPECTER2 embeddings have better precision values than VSR. Higher levels of recall are more relevant, and at these points, SPECTER2 embeddings have better retrieval accuracy. When evaluated with NDCG, SPECTER2 embeddings do worse than VSR, except for the Adapter model, which does better only at rank 1, with and without using cosine. Therefore in terms of NDCG, VSR has better retrieval accuracy.

2. Is there a difference with and without using the adapters?

The adapter model has higher NDCG values suggesting that the adapter-enhanced model is better at ranking the most relevant documents higher in the retrieval results, which would be useful in search tasks where ranking is critical. The adapter model also has lower precision-recall performance implying that while the adapter model ranks some relevant documents exceptionally well (as reflected in NDCG), it may miss others or retrieve some irrelevant documents, reducing overall precision and recall compared to the base model.

3. Why do you think classic VSR might still be out-performing these modern deep-learning methods (designed specifically for scientific documents) on this particular scientific corpus?

SPECTER2 generalizes the documents by encoding semantic relationships, but this might dilute the precision of retrieval for this particular scientific corpus. VSR uses the tf-idf weighting for the vector representations. Scientific documents usually have terminology that's highly specific, so the direct tf-idf approach could be more effective than abstract embeddings that deep-learning methods use. Additionally, SPECTER2 is trained on a large corpus of scientific documents, so the training data might not fully and accurately represent the terminology of this particular corpus.

4. How does using Euclidian distance vs. cosine similarity affect the results using the two deep models?

On the PR graphs, euclidean distance and cosine similarity don’t have an impact since they return similar relative ranking orders since embeddings are normalized, so there is negligible impact on PR results. But on the NDCG graph, cosine has slightly higher values. This is most likely because cosine measures the angle and emphasizes direction of vectors instead of magnitude. This is more beneficial in this context because the space has a very high dimension, so the angle probably has more indication of the semantic meaning than magnitude of the vectors do.

5. Does the hybrid solution improve over both methods. Why or why not?

The hybrid solution improves results when balancing the deep-learning and VSR methods appropriately. The effectiveness of the hybrid depends critically on the choice of lambda. As lambda increases, the hybrid depends more on the deep retrieval score. As mentioned earlier, the deep learning methods attempt to focus on semantic extraction, but sometimes overlook exact word matches that are central in scientific corpora, where specific jargon can be key. Overweighting the deep component with a higher lambda can dilute the impact of exact term matches that VSR picks up on, leading to poorer retrieval for highly specialized queries. Moderate values of lambda balance the semantic flexibility of deep learning and exact term matching of VSR and leverage the unique strength of both methods. Scientific datasets benefit from this balance because they require precision to retrieve papers with specific words and also semantic understanding to capture broader conceptual similarities.

6. What seems to be the best value of the hyperparameter λ?

The performance of the hybrid model seems to peak when lambda = 0.8. The PR and NDCG results are maximized at this value.