Project: The OGBN-MAG dataset is a subset of the Microsoft Academic Graph (MAG). The graph is a heterogeneous network with four types of entities: papers, authors, institutions, and fields of study. The network is composed of directed edges which connect two types of entities: an author is “affiliated with” an institution, an author “writes” a paper, a paper “cites” a paper, and a paper “has a topic of” a field of study. The prediction task for this benchmark is to predict the venue (conference/journal) of each paper, with there being 349 different venues.

Motivation: One strategy to improve predictive performance is to augment the heterogeneous network with additional nodes/edges that may provide useful knowledge during message aggregation. Data augmentation is a powerful method for training models that creates more robust representations, especially for sparse and imbalanced data. We investigate the effect that the numerosity of a paper’s fields of study connections have on prediction to determine if a paper within an “unpopular” field of study is more likely to be mispredicted.

Fields of Study Analysis: We ran five trials where each trial took a random sample of ten-thousand papers. From here we checked to see if this paper was predicted correctly or not, and then found the number of papers that the field of study it belongs to has. As seen in **table 1**, the results from the trial were inconclusive with the falsely predicted papers average being slightly higher.

Table 1. Papers Per Field of Study Analysis

	True		**False**	
Trial Number	Mean	Std Dev	Mean	Std Dev
#1	98,103	228,132	98,499	228,565
#2	97,970	227,790	98,513	228,489
#3	98,103	228,073	98,086	228,028
#4	98,017	227,883	98,137	228,139
#5	97,937	227,686	98,318	228,387

We saw that there is no statistically significant relationship between the incorrectly predicted papers and the “popularity” of the fields of study associated with it.

Cites Edge: We have an additional hypothesis that a paper’s future citations has an important effect on the prediction of one’s venue. Because future papers are likely to be related to the paper in terms of content, these papers should have a substantial impact on the venue prediction. Furthermore, we also found that papers that were correctly predicted had more authors, on average, than ones that were falsely predicted. We validate our hypothesis by masking the citation edges between papers.

Running the model with the same set of train, validation, and test data but with masking this edge yielded a test accuracy of 51.3198%, which is noticeably lower than the original model's 56.6895% accuracy. Furthermore, we ran the SeHGNN model with k-fold cross validation with ten folds in order to further validate our hypothesis. The results from this trial are shown below in Table 2.

Table 2. K-Fold Results

Fold Number	**Masking Test Accuracy (%)**	**No Masking Test Accuracy (%)**
1	56.9339	57.2422
2	56.8633	56.2332
3	56.4668	56.7438
4	56.2142	56.2842
5	54.6022	56.2842
6	53.5593	55.1284
7	54.1402	56.0182
8	53.4914	55.3827
9	54.4724	55.9204
10	53.1982	55.3829
Avg:	**54.99419**	**56.06202**

Running the SeHGNN model with ten k-folds shows that papers with their future citations masked tend to have a lower accuracy. Based on these results, we predict that augmenting the graph's edges for future citations will increase the model's accuracy.