GraphTranslate: Predicting Clinical Trial Translation using Graph Neural Networks on Biomedical Literature

Emily Muller and **Justin Boylan-Toomey** and **Jack Ekinsmyth**
Arne Robben and **María De La Paz Cardona** and **Antonia Langfelder**
Wellcome Trust
London UK

Abstract

The translation of basic science into clinical interventions represents a critical yet prolonged pathway in biomedical research, with significant implications for human health. While previous translation prediction approaches have focused on citation-based and metadata metrics or semantic analysis, the complex network structure of scientific knowledge remains under-explored. In this work, we present a novel graph neural network approach that leverages both semantic and structural information to predict which research publications will lead to clinical trials. Our model analyses a comprehensive dataset of 19 million publication nodes, using transformer-based title and abstract sentence embeddings within their citation network context. We demonstrate that our graph-based architecture, which employs attention mechanisms over local citation neighbourhoods, outperforms traditional convolutional approaches by effectively capturing knowledge flow patterns (F1 improvement of 4.5 and 3.5 percentage points for direct and indirect translation). Our metadata is carefully selected to eliminate potential biases from researcher-specific information, while maintaining predictive power through network structural features. Notably, our model achieves state-of-the-art performance using only content-based features, showing that language inherently captures many of the predictive features of translation. Through rigorous validation on a held-out time window (2021), we demonstrate generalisation across different biomedical domains and provide insights into early indicators of translational research potential. Our system offers immediate practical value for research funders, enabling evidence-based assessment of translational potential during grant review processes. The code for GraphTranslate is available at https://github.com/wellcometrust/graph-translate.

1 Introduction

The path from scientific discovery to clinical application remains a critical challenge in biomedical research. Although laboratory research and preclinical studies can lead to advances in scientific understanding, translating these findings into real-world clinical interventions that directly benefit patients is a complex process involving multiple stages, including experimental validation, regulatory approval, and clinical trials, each of which introduces uncertainty and challenges (Contopoulos-Ioannidis et al., 2008). Moreover, despite substantial global investment in medical research and development, only a tiny fraction of basic research findings successfully translate into clinical treatments (Contopoulos-Ioannidis et al., 2003). This inefficiency in the translation pipeline, combined with the decades-long timeframe typically required for bench-to-bedside translation (Morris et al., 2011), creates an urgent need for tools that can identify promising translational research early in its lifecycle.

Previous approaches to predicting translational success have primarily relied on citation patterns and metadata features, or focused solely on semantic analysis of research content (Nelson et al., 2022; Padilla-Cabello et al., 2022). While these methods have shown promise, they often overlook the complex network of scientific knowledge through which research findings propagate towards clinical applications. Citation networks represent more than just academic impact—they capture the flow of knowledge from fundamental discoveries towards clinical implementation. However, effectively modelling these knowledge transmission pathways requires both understanding the semantic content of research and its structural position within the broader scientific landscape.

We address this challenge by introducing a graph neural network architecture that integrates

Proceedings of the Fifth Workshop on Scholarly Document Processing (SDP 2025), pages 31–41

both semantic and structural information from research publications. By analysing a comprehensive dataset of 19 million publications using transformer-based embeddings within their citation network context, our model captures subtle patterns in how knowledge flows from basic science towards clinical applications. Crucially, we demonstrate that content-based features inherently encode many signals predictive of translational potential, allowing us to achieve state-of-the-art performance while minimising reliance on potentially biased metadata features. As a result, this approach offers practical value in identifying promising translational research early, helping researchers, funders, and institutions prioritise high-impact projects.

2 Related Work

Recent academic enquiry has focused on predicting the relationship between a paper and its translational outcomes via citation analysis. Hutchins et al. discovered a complex relationship between a paper's content, its citing articles, and citation rates, affecting its likelihood of being cited in clinical articles (Hutchins et al., 2019b). The study used human-annotated Medical Subject Headings (MeSH) and 22 features in a random forest model to predict translational success. The model achieved good accuracy with just two years of data, and the authors showed diminishing improvement when more years of data were added. This is crucial because early identification facilitates translation prediction within the timeframe of a grant.

The predictive power of MeSH tags is attributed to its identification of the clinical stage a paper lies (Hutchins et al., 2019b). The use of MeSH terms, however, is limiting, as it requires extensive human labelling. As an alternative, modern natural language processing methods can also identify the translational stage of a paper (Li et al., 2023). Full-MLP-CNN model predicted patent citations (AUROC = 0.915) and guidelines and policy documents (AUROC = 0.918) without MeSH terms (Nelson et al., 2022). This model used sentence embeddings of the title and abstract alongside extensive metadata from the Microsoft Academic Graph (MAG). This included Microsoft's ranking features based on eigencentrality. These features assess the scientific network surrounding entities such as papers, journals, or authors, with the rank of the paper emerging as the most influential metadata feature.

The Full-MLP-CNN study did not address how the age of the paper affects the accuracy of the prediction. This is important as, when time limited, more complex network measures can perform poorly compared to citation counts (Mariani et al., 2016). However, Microsoft explicitly sought to mitigate this age bias in their entity centrality metrics via reinforcement learning (Wang et al., 2019). In fact, a time-balanced network centrality measure has been shown to be more effective than simple citation counts in identifying Nobel Prize winning papers even in the first few years (Mariani et al., 2016). This indicates that even a time-limited citation network structure contains valuable information for translation prediction. The DELPHI model is a clear example of this (Weis and Jacobson, 2021). It combined article and journal metadata alongside a time-limited citation network to predict 5-year post-publication time-balanced network centrality using only 2 years of data, while identifying seminal biotechnology papers.

Nelson et al. found that removing citation metadata features had a moderate reduction in predictive performance. This is supported by Li et al., who expanded on the NIH study using a total of 91 citation and MeSH based features to predict the clinical citation count of papers (Li et al., 2022). The authors found the expanded citation network from paper references were more influential than those of the predicted paper or its early citations (Xin Li, Xuli Tang and Qukai Cheng, 2022). Beinat et al. meanwhile, showed within the fields of dementia and cancer, translation could be predicted without any citation network data (Beinat et al., 2024). They used an array of article metadata features alongside title and abstract embeddings in a CatBoost model to predict patent (AUROC = 0.84) and clinical trial citations (AUROC = 0.81) for dementia research.

Removing citation features is attractive, as it allows for translation prediction without any time delay. However, models from Nelson et al., and Beinat et al. use an array of researcher features in their models, raising concerns about increased model bias. While author popularity may relate to past translational success, it is difficult to separate this from potential structural biases when predicting future success. We argue that assessing a paper's translational potential should not consider personal researchers attributes such as h-index, institution or country.

Evidence suggests that both paper content and a

time-limited scientific network structure around papers can be used to effectively predict biomedical translation. However, no study to date has successfully integrated both elements. Our graph neural network approach combines paper embeddings alongside a time-limited scientific network structure to achieve this. The final model successfully predicts translational impact without depending on extensive feature engineering (MeSH), a now discontinued service (MAG), and minimises bias by excluding sensitive author features.

3 Methodology

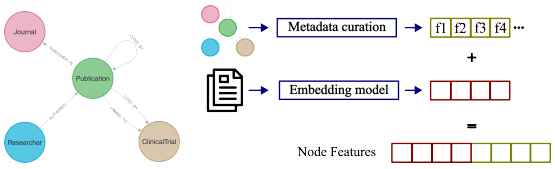
3.1 Data

3.1.1 Wellcome Academic Graph

Our dataset was extracted from our custom-built graph database deployed on AWS Neptune, the Wellcome Academic Gaph (WAG). WAG is a network model of the academic landscape, with nodes representing academic entities and edges representing interactions between these entities. It is modeled on the retired Microsoft Academic Graph (Sinha et al., 2015) and tailored to meet our organisation's analysis requirements, including but not limited to enhanced coverage of grant funding data. WAG currently contains over 346 million academic entities (covering scientific publications dating back to the year 1665), connected by 2.9 billion edges. The underlying source data are based on Dimensions (Digital Science) (Herzog et al., 2020), a commercially available scientific research database commonly used by research funders, which we augmented with internal data. The latest version of WAG includes an enrichment layer to add precomputed metrics and relationships to the graph. Figure 1a shows part of the graph schema relevant for GraphTranslate. Dimensions covers a wide range of articles from open- and closed-access journals (Singh et al., 2021), as well as clinical trial records from 15 registries (as of 2023) (Resources, 2018). Instead of citations from clinical trial publications, we use citations provided as part of these clinical trial records as the target label to predict translation.

3.1.2 Preprocessing

The data were filtered to include only publications related to medical science, as defined by the ANZSRC Field of Research (FoR) codes from 2020 (of Statistics, 2020), which are provided as part of the source data of the publication. The following



(a) Graph schema (b) Publication node features

Figure 1: Academic graph database schema and the construction of publication node embeddings.

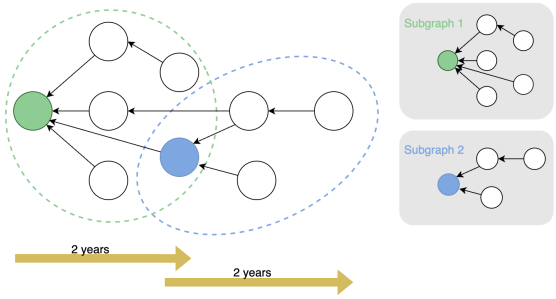


Figure 2: Citation data loading.

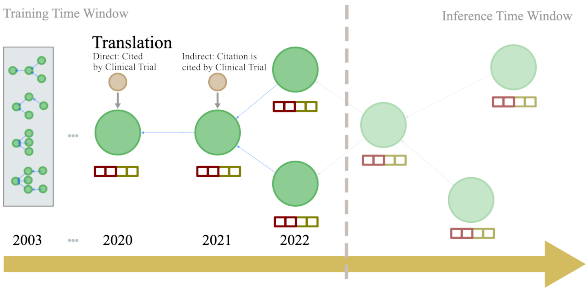


Figure 3: Temporal diagram of translation prediction.

Division-level FoR codes were selected based on our own exploratory analysis of publications historically cited by clinical trials: Biomedical and Clinical Sciences (32), Health Sciences (42), and Psychology (52). In addition, we limited our dataset to research articles by filtering on article-type tags. This was done to ensure that any performance metrics of the resulting models are both realistic (by excluding articles which will conceivably never be cited by a clinical trial) and indicative of translational potential of original research (by excluding review articles, among others). The publications' local citation network was extracted within a 2-year time window used for graph modeling, a time period previously identified as sufficient for predicting citation by a clinical article (Hutchins et al., 2019b). As shown in Figure 2, this was done by loading each year's publication nodes together with citations covering their respective 2-year time win-

dow as distinct sub-graphs.

3.1.3 Text Embeddings

Given our focus on semantic information as a key node feature for prediction, we included only those articles which had English language titles and abstracts available. Non-English texts were filtered out using the Google Compact shallow language detector network (Google). Semantic node features were created by converting titles and abstracts to text embeddings using SciBERT, a language model pre-trained on a multi-domain corpus of scientific publications, released in 2019 (Beltagy et al., 2019). Titles and abstracts were concatenated and tokenized. To produce fixed-length embeddings, longer texts were truncated while shorter texts were padded to the maximum sequence length of 512 tokens. 768-dimensional representations of titles and abstracts were produced by applying mean-pooling to the token-level embeddings generated by the model.

3.1.4 Graph Loader

We preprocessed the citation network by filtering out publications that received no citations (approximately 51% of the dataset) and those lacking required metadata fields. The final dataset was split into training (80%), validation (10%), and test (10%) sets for each year.

For training efficiency and to address class imbalance (1.7% positive cases), we downsampled the majority class in the training set to achieve a 1:1 ratio. The validation and test sets maintain their original class distributions to reflect real-world conditions. We implemented a custom PyTorch Geometric DataLoader with a batch size of 256 to handle the large-scale graph structure, using neighbor sampling with a maximum of 500 nodes in the first layer and 1000 nodes in the second to manage memory constraints.

3.2 GNN Model

3.2.1 Model architecture

We implemented a binary node classifier using a graph neural network (GNN) approach. GNNs are typically built on the assumption that the input graph is undirected. However, we hypothesised that our citation network's inherent directionality carries predictive value indicative of translation. Specifically, we hypothesized that data from publications cited by the article in question (i.e., past publications) are less informative than the citations an article receives after its publication. To leverage this directional information in our model, we used a Directed Graph Neural Network (DirGNN) architecture as first described by Rossi et al. (Rossi et al., 2023). We considered 2-hop citation neighborhoods when updating our node features for translational prediction, implemented using two graph convolutional layers for message passing. We compared the following graph convolutional layers, which are available as part of Pytorch Geometric: a simple graph convolutional operator (GCNConv) (Kipf and Welling, 2017), a GraphSAGE operator (SAGEConv) (Hamilton et al., 2018), and a graph attentional operator (GATConv) (Veličković et al., 2018). We applied the ReLU activation function and dropout after each convolutional layer. Our best performing model was trained using two GATConv layers with 32 hidden dimensions. Furthermore, we implemented jumping knowledge layer aggregation as part of our GNN architecture, which was based on concatenation of the model's hidden representations (Xu et al., 2018). Finally, a linear output layer was used to generate logits for binary classification.

3.2.2 Model training

Our GNN model was implemented in PyTorch Geometric. We used the Adam optimiser with a learning rate of 1e-3. Models were trained for 50 epochs with early stopping based on validation loss with a patience of 5 epochs. The hidden dimension was set to 32 with 2 graph attention layers. Dropout of 0.2 was applied after each layer. Model training was performed on a cloud compute instance with a Nvidia A10G GPU. Hyperparameters were determined through systematic grid search optimisation.

4 Experiments

To evaluate the efficacy of our graph-based approach for predicting research translation into clinical trials, we conducted four sets of experiments designed to test key hypotheses about model performance, feature importance, early detection capabilities and comparison to previous literature.

4.1 Graph Neural Networks vs. Linear Baseline

Our first experiment compares the predictive performance of our graph-based approach against a traditional linear baseline. Both models were trained

on identical datasets comprising academic publications and their associated clinical trial citations. The baseline architecture consists of three linear layers (64 units each) with dropout regularisation (p=0.1). Our proposed graph model implemented two Graph Attention layers (GATConv) with 32-dimensional hidden representations and dropout (p=0.2). For comprehensive evaluation, we considered both direct (publications cited by clinical trial) and indirect (publications' citation is cited by clinical trial) connections between publications and clinical trials in our network structure.

4.2 Publication Node Metadata

We evaluated model performance with different types of metadata features: citation count, Field of Research classifications (FoR), Research Activity Classifications (RAC) (, UKCRC), journal impact metrics, and historical clinical trial participation by any authors. FOR classifications provide broad labeling of fields such as Biological Sciences (top-level class) and Ecology (second-level class). RAC classifications are specific to health-related research with 48 distinct codes organised into eight overarching groups. We obtained historic journal metrics data from Scimago API for each publication (Scimago, 2024). Historic clinical trial participation of an author required that any author be previously associated with a publication directly linked to a clinical trial (not cited). To manage high-dimensional feature spaces (>32 dimensions), we applied Principal Component Analysis (PCA) and retain the top 32 principal components. These reduced metadata embeddings are concatenated with the document text embeddings before being passed through the network.

4.3 Early Detection Performance

To assess the model's capability for early identification of translational research potential, we evaluated direct translation prediction on recent publications in the inference time window (2021). This experiment particularly focuses on the model's ability to identify longer pathways of translational research early on.

4.4 Evaluation

We evaluated the performance of our direct model on NIH's publicly available dataset: iCite (Hutchins et al., 2019a). We removed the feature which links authors to previous clinical trials and retrained our model for inference on this dataset using the same hyperparameters and early stopping. After removing publications without any citations or embeddings, there are a total of 5 million publications between 2003 and 2020. There is a 50% translational rate in this dataset.

Table 1: Validation dataset performance comparison between Linear Model and Graph Neural Network.

Model	AUROC	Recall	Precision	F1	AP
Dir. Linear Model	0.786	0.653	0.088	0.155	0.092
Dir. GNN	0.831	0.647	0.120	0.203	0.132
Indir. Linear Model	0.783	0.390	0.675	0.494	0.532
Indir. GNN	0.818	0.647	0.618	0.632	0.596

5 Results

5.1 Citations

Our dataset comprises 19 million publications and 127 million citation edges from 2003 to 2020 within the training window. Among these publications, 1.7% were identified as directly translational and 14.3% identified as indirectly translational. Analysis reveals average translation times of 6 ± 4 years for a clinical trial citation. The inference window (2021) contains 1.4 million publications, with 0.6% identified as translational. Publications are labeled as translational if they have been cited by a clinical trial as of April 2024. Evaluation of predictions using post-April 2024 clinical trial citations are reported for the test performance.

5.2 Baseline Model Performance Comparison

Our graph neural network (GNN) architectures demonstrate superior performance compared to baseline linear models across both direct and indirect translation prediction tasks. Both GNN models, which incorporate only embedding-based node attributes, effectively capture not only the semantic context of the target publication but also the structural information from its citation neighborhoods. This dual representation leads to improved overall performance metrics, with the direct translation GNN achieving an F1 score improvement of 4.5 percentage points (0.155 vs 0.203) and average precision increase of 4 percentage points. Similarly, the indirect translation GNN demonstrates an F1 score improvement of 12.8 percentage points (0.494 vs 0.632) and average precision of 4% over its linear counterpart.

5.3 Impact of Node Metadata Features

Analysis of different node metadata features reveals varying contributions to model performance. Re-

Table 2: Validation dataset performance comparison for metadata features.

Metadata	AUROC	Recall	Precision	F1	AP
Embeddings	0.831	0.647	0.120	0.203	0.132
+Cite	0.846	0.751	0.105	0.184	0.132
+Cite+Journal	0.722	0.337	0.136	0.194	0.072
+Cite+FOR	0.845	0.794	0.095	0.170	0.135
+Cite+RAC	0.831	0.685	0.130	0.218	0.151
+Cite+FOR+RAC	0.834	0.772	0.110	0.192	0.144
+Cite+Prev.Trial	**0.855**	**0.802**	**0.101**	**0.179**	**0.138**
+Cite+Prev.Trial+FOR	0.855	0.797	0.100	0.178	0.137
+Cite+Prev.Trial+RAC	0.849	0.666	0.145	0.239	0.161

Table 3: Test dataset performance for direct and indirect metadata models.

Model	AUROC	Recall	Precision	F1	AP
Direct GNN	0.852	0.788	0.107	0.188	0.148
Indirect GNN	0.815	0.551	0.662	0.601	0.551

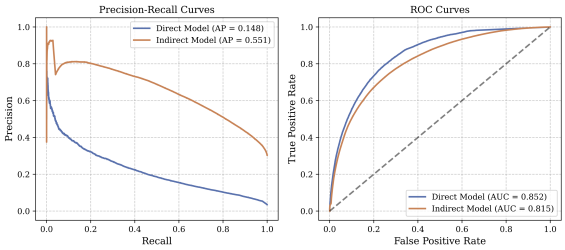


Figure 4: Precision-Recall and ROC curves for direct and indirect test performance.

search Activity Classification (RAC) codes provide the strongest performance boost, increasing the average precision (AP) to 0.151. These codes, which specifically categorise health and clinical research domains, offer an additional health-oriented perspective for assessing translational potential. However, their impact is limited by sparse coverage, with only 13% of publications having RAC annotations.

Fields of Research (FOR) codes, despite their broader coverage across the citation network (>80%), do not significantly improve either F1 or AP scores. Author-based features derived from previous clinical trial associations demonstrate the second-highest performance improvement (AP increase of 0.138), representing the marginal increased gain in scenarios where RAC codes are unavailable. Contrary to guideline/policy and patent prediction (Nelson et al., 2022), the inclusion of journal-based metrics (e.g., impact factor, citation counts) degraded model performance, suggesting that traditional bibliometric measures may not be reliable indicators of translational potential.

5.4 Test Performance

The final direct and indirect models trained on the embeddings, citation count and researcher linked clinical trials were used to measure test performance metrics as shown in Table 3, and precision-recall and ROC curves, as shown in Figure 4. The indirect model is able to substantially improve the precision on the test dataset. This is owing to a more balanced dataset with 14.3% of publications being indirectly cited by a trial.

A closer inspection of the performance metrics across the years show that the accuracy metrics are non-stationary over time, with more recent years suffering a degradation (see Appendix Figure 6). This is most likely due to the recency of these publications and the limited time elapsed to complete clinical trial citations compared to previous years. The assumption, therefore, is that a proportion of the false positives are incorrectly labeled as such. In order to validate this, we collected new Clinical Trials (as at January 2025). Analysis of Clinical Trials data post April 2024 reveals 5,421 new trials linked to 48,021 historic publications. Of these newly translated publications, 1,373 intersect with our test set. When accounting for these recent trials, 0.5% of all test publications were initially mislabeled as false positives. The updated precision scores per year are shown in Appendix Figure 7, with more recent years having a greater proportion of incorrectly labeled false positives. This indicates our model's ability to identify publications with future translational potential.

Appendix Figure 8 demonstrates varying performance for different fields of research. For health sub-domains: neuroscience, reproductive medicine, health and clinical sciences have the highest precision scores (above the global average). These fields represent scientific domains which may often include animal or human participants, positioning them closer to translational outcomes. On the other hand, the health fields with lower precision include biological sciences, medical biotechnology, engineering and microbiology. While a proportion of this can be attributed to longer translation pathways, certain fields continue to demonstrate increased performance pre-2010 (see Appendix Figure 9 - clinical sciences increases by 4 percentage points compared to biological sciences which increased by 1 percentage point). This indicates that the model is better at identifying translation in certain biomedical domains using the publication text and network neighbourhood.

Table 4: Inference performance for direct model.

Model	AUROC	Recall	Precision	F1	AP
Direct GNN	0.728	0.327	0.135	0.191	0.080

5.5 Early Detection Performance

Recall that on average it takes approximately 6 years to obtain a citation from a clinical trial (based on the Wellcome Academic Graph data). Since the inference time window includes publications from 2021, publications have accrued only 3.25 of post-publication citations, resulting in incomplete ground labels, as indicated by a low citation rate in the inference time window (0.6% versus 1.7%). This results in a degradation of the model recall as shown in Table 4. We would expect these results to be recovered once a more complete time window has elapsed.

The translation model predicts the field of immunology to have the greatest number of translational publications in 2021. This is followed by epidemiology and medical microbiology. These are predicted to translate at a rate of close to 4%. However, the precision score is likely to reduce that rate by a factor of 10 for true translation proportion. These fields reflect the translational contribution towards understanding Covid-19 during the pandemic (in the test dataset immunology had a translation rate of 2%).

We updated Clinical Trial citations by importing data after April 2024 up until January 2025. This led to incorrect false positives for the inference time window to have a 1.4% error rate. The correctly predicted publications have a very high median citation count (median > 100), indicative of high impact translational research (see Appendix Figure 11). In contrast the set of false negatives have a much lower citation count (median ≈ 10). The Research Activity Codes (RAC) associated with higher proportion of false negatives (see Appendix Figure 12) include individual care needs, organisation and delivery of services and primary prevention interventions to modify behaviours or promote wellbeing. These fields represent research close to translational science, and as shown by (Li et al., 2024), can have a low number of overall non-clinical citations. Lower citation count is not only a feature used for prediction, but also reduces the aggregated network neighbourhood effects for the GNN model prediction.

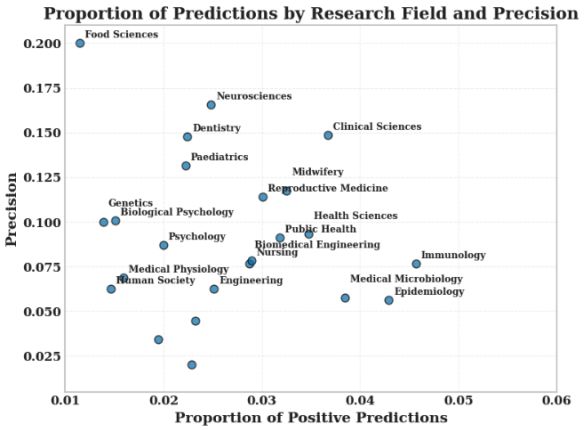


Figure 5: Proportion of positive predictions per field of research versus precision.

Table 5: Inference performance for NIH iCite dataset.

Source	AUROC	Accuracy	F1
Direct GNN	0.85	0.78	0.64
NIH RF (Hutchins et al., 2019b)	0.80	0.84	0.56

5.6 iCite Evaluation

Table 5 shows the performance of our model on the sampled iCite dataset. For completeness, we also show the reported results from the original model which incorporated MeSH categories of the original publication and its' first-order citations. The original publication reports a translation rate of 30% compared to our 50% (Hutchins et al., 2019b). This is most likely due to the time frame: the authors use publications from 1994 - 2014. Our model outperforms the NIH model on the iCite dataset using only embedding and citation count as features. This demonstrates the ability of title and abstract embeddings to encode information on a meaningful scale for understanding clinical trial translation. Since our model does not need any research codes such as MeSH, it can be applied to a broader set of research publications outside of the PubMeD archive.

5.7 Discussion

Our work presents the first application of graph neural networks to model research translation by incorporating local citation network structures. Following (Hutchins et al., 2019b), we use a two-year post-publication citation window, optimising for early detection while maintaining predictive power. However, our approach differs significantly in dataset scope - while their study focused exclusively on PubMed-indexed publications, we analyse a broader spectrum of biological and health

sciences research, resulting in a more realistic but challenging 1.7% translation rate. This wider scope leads to lower F1 scores (0.19 versus 0.56), reflecting the inherent difficulty of prediction in a more imbalanced, real-world setting. We find that our model outperforms the model from NIH when predicting on the NIH-iCite dataset leading to increased AUC and F1 scores (0.64 versus 0.56 and 0.85 versus 0.80).

Our results demonstrate that graph neural networks, particularly through attention mechanisms, more effectively capture translation patterns compared to linear combinations of node, MeSH and citation features. The inclusion of Research Activity Classification (RAC) codes provides the highest performance boost, potentially serving a similar role to the MeSH-based features in Hutchins et al. (Hutchins et al., 2019b). However, the marginal improvement from RAC codes suggests our node embeddings already encode much of this information. Additionally, MeSH terms only cover a subset of publications (associated with PubMeD), therefore an embedding based approach to quantifying research content allows for greater generalisability. Notably, contrary to Nelson et al. (Nelson et al., 2022), we found no performance improvement from journal metrics, though this may reflect our focus on clinical trials rather than guidelines and policy citations.

In an early prediction time window, our model maintains precision capabilities while experiencing expected recall reduction due to the time lag between publication and clinical trial citation. Correct predictions correlate with early citation impact, while false positives concentrate in fields with typically longer translation pathways, such as biological mechanisms and oncology research. This pattern suggests our model effectively captures domain-specific translation dynamics.

In future work, a time-normalised PageRank measurement could improve model performance without relying directly on citations. Unlike (Nelson et al., 2022), we deliberately excluded potentially biased features such as researcher demographic or impact scores to avoid potential bias from academics with a longer career history or from certain well funded-geographies. Model performance might be further improved by incorporating local citation networks (references) to provide additional insights into knowledge flow patterns, an approach that has shown promise in predicting clinical citation patterns (Li et al., 2022).

Limitations

Our study has several important limitations that should be considered when interpreting results and applying the model:

Temporal Constraints

The model requires at least 2 years of citation data post-publication to make reliable predictions. This creates an inherent delay in assessment capabilities, limiting its use for real-time funding decisions. Since research grants themselves require time to produce outputs, funders would need to wait a minimum of 2 years from grant award date to assess translation potential. Consequently, this approach is more suitable for retrospective analysis of research portfolios where manual labeling would be impractical or unfeasible.

Limited Translation Outcomes

Our current implementation defines translation narrowly through citation in clinical trials. This definition excludes other important translation pathways such as patents, policy documents, clinical guidelines, commercial products, and public health interventions. In addition, it does not differentiate between phases of Clinical Trials. A more comprehensive model would incorporate these diverse translation outcomes to better reflect the multifaceted nature of research impact beyond the clinical trial pathway.

Interpretability Challenges

The complex interactions captured by graph neural networks limit straightforward interpretation of why specific research is predicted to translate. This "black box" aspect may limit acceptance by funding bodies or policymakers who require transparent decision-making rationale.

Ethics Statement

As authors submitting research to the ACL conference, we affirm our commitment to ethical standards through honest reporting, accurate data, and rigorous scientific methods. Our work provides transparent details for reproducibility, acknowledging ethical considerations, particularly regarding privacy, fairness, and potential misuse. We've respected copyrights and intellectual property, obtaining necessary permissions and crediting sources.

Our research promotes diversity, inclusion, and equity, deliberately avoiding biases related to gender, race, ethnicity, or any other characteristic. Compliance with ethical guidelines, including those by ACL and our institutions, remains paramount. By submitting this paper, we assert adherence to these standards and pledge to address any ethical concerns arising during the review.

References

Matilda Beinat, Julian Beinat, Mohammed Shoaib, and Jorge Gomez Magenti. 2024. Machine learning to promote translational research: predicting patent and clinical trial inclusion in dementia research. *Brain Communications*, 6(4):fcae230.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*.

D. G. Contopoulos-Ioannidis, G. A. Alexiou, T. C. Gouvias, and J. P. Ioannidis. 2008. Life cycle of translational research for medical interventions. *Science*, 321(5894):1298–1299.

Despina G Contopoulos-Ioannidis, Evangelia Ntzani, and JP Ioannidis. 2003. Translation of highly promising basic science research into clinical applications. *The American journal of medicine*, 114(6):477–484.

Google. Compact language detector v3 (cld3).

William L. Hamilton, Rex Ying, and Jure Leskovec. 2018. Inductive representation learning on large graphs.

Christian Herzog, Daniel Hook, and Stacy Konkiel. 2020. Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, 1(1):387–395.

B Ian Hutchins, Kirk L Baker, Matthew T Davis, Mario A Diwersy, Ehsanul Haque, Robert M Harriman, Travis A Hoppe, Stephen A Leicht, Payam Meyer, and George M Santangelo. 2019a. The nih open citation collection: A public access, broad coverage resource. *PLoS biology*, 17(10):e3000385.

B Ian Hutchins, Matthew T Davis, Rebecca A Meseroll, and George M Santangelo. 2019b. Predicting translational progress in biomedical research. *PLoS biology*, 17(10):e3000416.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks.

Xin Li, Xuli Tang, and Qikai Cheng. 2022. Predicting the clinical citation count of biomedical papers using multilayer perceptron neural network. *Journal of Informetrics*, 16(4):101333.

Xin Li, Xuli Tang, and Wei Lu. 2023. Tracking biomedical articles along the translational continuum: a measure based on biomedical knowledge representation. *Scientometrics*, 128(2):1295–1319.

Xin Li, Xuli Tang, and Wei Lu. 2024. How biomedical papers accumulated their clinical citations: a large-scale retrospective analysis based on pubmed. *Scientometrics*, 129(6):3315–3339.

Manuel Sebastian Mariani, Matúš Medo, and Yi-Cheng Zhang. 2016. Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics*, 10(4):1207–1223.

Zoë Slote Morris, Steven Wooding, and Jonathan Grant. 2011. The answer is 17 years, what is the question: understanding time lags in translational research. *Journal of the royal society of medicine*, 104(12):510–520.

Amy PK Nelson, Robert J Gray, James K Ruffle, Henry C Watkins, Daniel Herron, Nick Sorros, Danil Mikhailov, M Jorge Cardoso, Sebastien Ourselin, Nick McNally, et al. 2022. Deep forecasting of translational impact in medical research. *Patterns*, 3(5).

Australian Bureau of Statistics. 2020. Australian and new zealand standard research classification (anzsrc).

Javier Padilla-Cabello, Antonio Santisteban-Espejo, Ruben Heradio, Manuel J Cobo, Miguel A Martin-Piedra, and Jose A Moral-Munoz. 2022. Methods for identifying biomedical translation: A systematic review. *American Journal of Translational Research*, 14(4):2697.

Dimensions Resources. 2018. A Guide to the Dimensions Data Approach.

Emanuele Rossi, Bertrand Charpentier, Francesco Di Giovanni, Fabrizio Frasca, Stephan Günnemann, and Michael Bronstein. 2023. Edge directionality improves learning on heterophilic graphs.

Scimago. 2024. Journal and country rank.

Vivek Kumar Singh, Prashasti Singh, Mousumi Karmakar, Jacqueline Leta, and Philipp Mayr. 2021. The journal coverage of web of science, scopus and dimensions: A comparative analysis. *Scientometrics*, 126(6):5113–5142.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th international conference on world wide web*, pages 243–246.

UK Clinical Research Collaboration (UKCRC). 2024. Research activity code health research classification system (rac hrcs).

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks.

Kuansan Wang, Iris Shen, Charles Huang, Chieh-Han Wu, Darrin Eide, Yuxiao Dong, Junjie Qian, Anshul Kanakia, Alvin Chen, and Rick Rogahn. 2019. A review of microsoft academic services for science of science studies. *Frontiers in Big Data*, 2:45.

James W Weis and Joseph M Jacobson. 2021. Learning on knowledge graph dynamics provides an early warning of impactful research. *Nature Biotechnology*, 39(10):1300–1307.

Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken ichi Kawarabayashi, and Stefanie Jegelka. 2018. Representation learning on graphs with jumping knowledge networks.

A Appendix

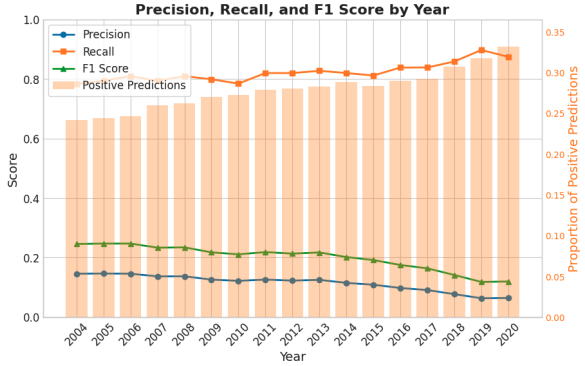


Figure 6: Test set performance metrics for direct clinical trial translation prediction.

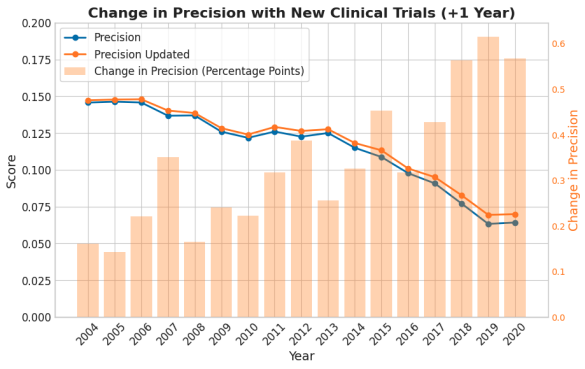


Figure 7: Test set precision metrics including updated metrics for newly translated publications.

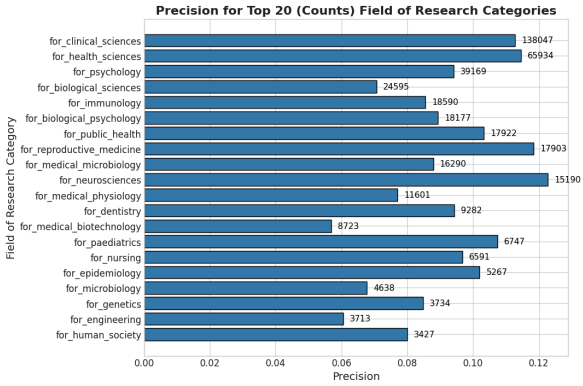


Figure 8: Test set precision metrics for most commonly appearing fields of research (total test set counts shown in brackets).

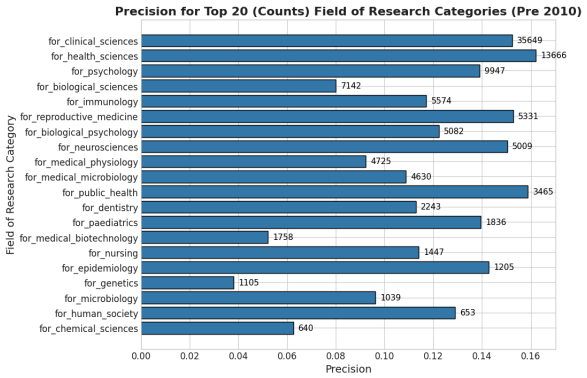


Figure 9: Test set precision metrics for most commonly appearing fields of research (pre-2010 only).

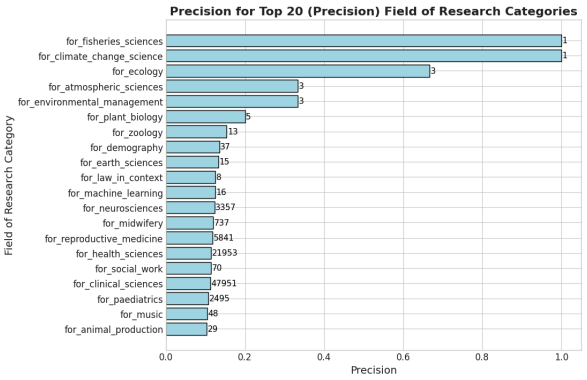


Figure 10: Test set precision metrics highest performing fields (total positive number of publicatoins shown in brackets).

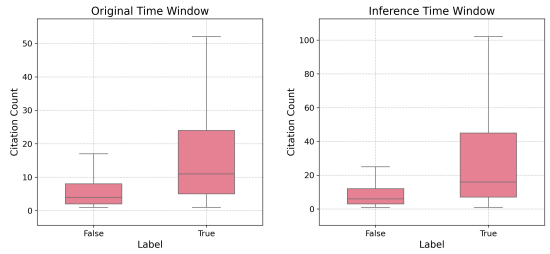


Figure 11: Distribution of citation count in training (left) and inference time window.

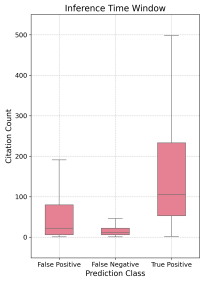


Figure 12: Citation count distribution in each inference prediction class.

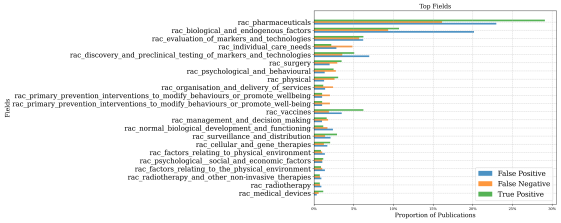


Figure 13: Proportion of RAC research fields in each inference prediction class.