Graph Foundation Model Biomedical Link Prediction Performance On An Evolving Knowledge Graph

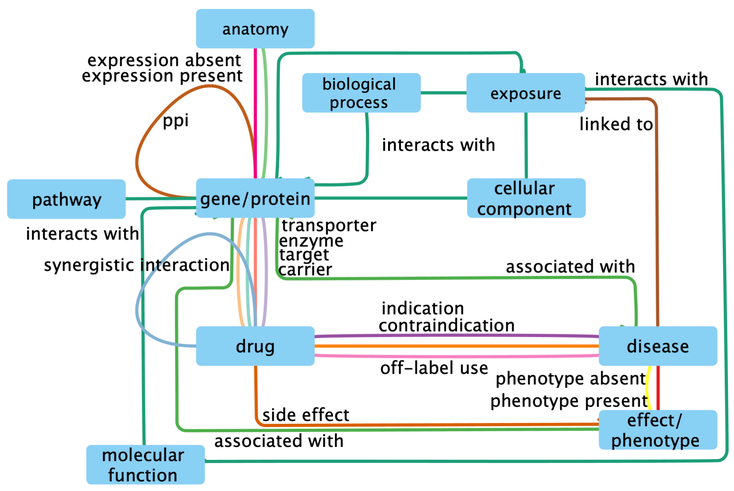
Roger Tu [1, 2], Daniel Myung [2], Dragan Djordjevic [2]

[1] Rancho Biosciences, Rancho Santa Fe, CA, USA; [2] Takeda Development Center Americas, Inc., Cambridge, MA, USA

Overview

- Biomedical Knowledge Graphs (KGs) are networks that describe concepts like drugs, genes, & diseases & their relationships (e.g., treats & interacts with). These networks are composed of triples (facts); concepts linked by a relation (e.g., atorvastatin treats hyperlipidemia).
- Query Answering (QA) approaches on KGs aim to discover missing graph facts. These models first learn "barcodes" for each entity & relation, which are then used by a model to make deterministic predictions. However, learned "barcodes" are idiosyncratic to the model & graph it was trained on. "Barcodes" must be relearned if the KG is updated.
- Like language models, KG foundation models (KGFM) learn features across diverse domains to be broadly generalizable. These models can address costly and time intensive retraining.
- In this work, a KGFM was benchmarked against classical KG embedding (KGE) models & the QA performance was measured as the KG evolved. Both finetuned & zero shot KGFM outperformed finetuned KGE approaches. However, as the graph evolves, further KGFM finetuning should be considered.

Figure 1. Knowledge Graph Schema



PrimeKG comprises of 8.1 million facts with each entity classified as one of10 biological types interconnected with 18 unique relationships

Methods

Knowledge Graphs

- The 2023 PrimeKG biomedical KG integrates 20 distinct resources & describes 8.1 million triples between 129,375 entities (**Figure 1**).
- Initial KG embedding & foundation model evaluations were run on PrimeKG; triples were split 8/1/1 into train, test & validation sets.
- The 2026 OptimusKG biomedical KG contains 21.8 million triples with 190,939 entities. 3.2 million new triples in OptimusKG were aligned to PrimeKG by filtering for shared nodes & relations.
- Five additional datasets were created to evaluate KG model QA performance on unseen facts (OptimusPrimeKG). All PrimeKG triples were allocated to train sets in each. An increasing concentration of new triples were then added to train, with the remaining triples split 1/1 into test & validation sets (**Table 1**).
- Graph connectivity was evaluated using geodesic distance, the shortest path between two entities. Eccentricity is the set of shortest paths in a graph.

Table 1. OptimusPrimeKG Titration Triple Counts

Condition	Train	Test	Validation
OptimusKG Holdout (0%)	8,100,498	1,555,983	1,555,982
PrimeKG + OptimusKG (20%)	8,722,891	1,244,786	1,244,786
PrimeKG + OptimusKG (40%)	9,345,284	933,590	933,589
PrimeKG + OptimusKG (60%)	9,967,677	622,393	622,393
PrimeKG + OptimusKG (80%)	10,590,070	311,197	311,196

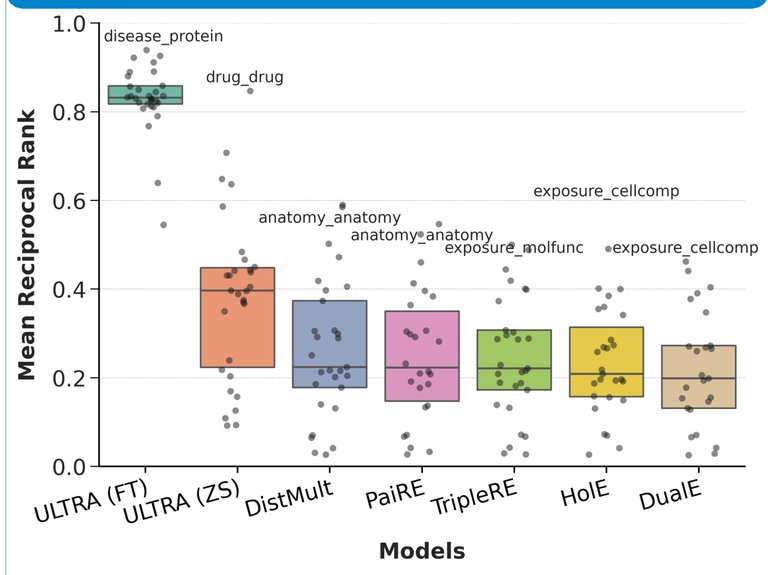
Five datasets were designed to measure the KG foundation model's predictive capacity on blinded facts. The train set is composed of aligned triples added to PrimeKG at a specified concentration; the remaining aligned triples were allocated to test & validation sets.

Model Finetuning & Evaluation

- ULTRA is a graph foundation model pretrained on 50 graphs across various knowledge domains. It learns invariances of graph structure to construct relative relation- and entity-representations (embeddings).
- ULTRA was finetuned on PrimeKG & each OptimusPrimeKG condition. The zero shot, PrimeKG finetuned & individual OptimusPrimeKG finetuned checkpoints were compared across conditions. All models were trained for an epoch using default training parameters with 8 x A100s (80GB) on AWS.
- Mean Rank, the average rank of the correct answer, & Mean Reciprocal Rank (MRR), the average sum of the reciprocal correct answer ranks, were measured.

Results

Figure 2. PrimeKG Model Query Answering Performance

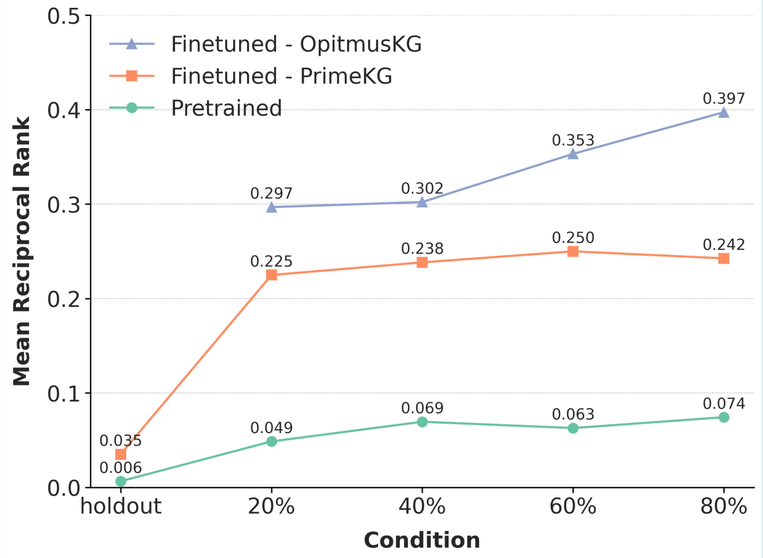


Evaluating edge type Mean Reciprocal Rank (MRR) distribution of pretrained- (orange, ZS) and finetuned- (teal, FT) foundation model against classical KG embedding models on PrimeKG. All models shown have been finetuned on the publicly available PrimeKG. The max edge MRR is labeled.

Static Graph Query Answering Evaluation

- The ULTRA foundation model was applied to PrimeKG to evaluate zero shot (ZS) performance. The finetuning (FT) & ZS MRR was 0.87581 & 0.63791, respectively. The edge specific MRR increased substantially after FT compared to ZS (**Figure 2**).
- ULTRA's edge specific query answering ability was compared to published results of PrimeKG finetuned KG embedding models. Edge specific MRR are derived by aggregating triples into triple types; this measure allows triple type specific evaluation. ULTRA's pretrained and finetuned models edge specific MRR mostly outperformed the traditional KG embedding models.

Figure 3. OptimusPrimeKG Titration Model Evaluation

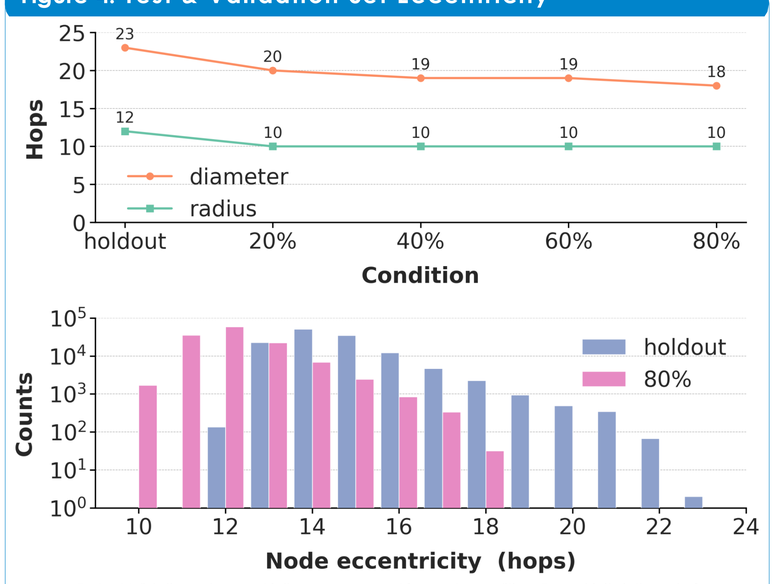


Model performance on OptimusKG edges. Holdout has each model predict 100% OptimusKG edges using PrimeKG as the training set. Both Pretrained and Finetuned – PrimeKG checkpoints were run in a ZS configuration. Finetuned – OptimusKG was Finetuned on each condition. Finetuning improves MRR.

Titration Graph Query Answering Evaluation

- ULTRA's ZS & PrimeKG FT query answering performance was measured on OptimusKG holdout edges (**Figure 3**, holdout). Both performed poorly; the PrimeKG FT was marginally better than ZS checkpoint with MRR of 0.0349 & 0.0065, respectively.
- Both ULTRA's ZS & PrimeKG FT performance increased as the train subgraph accumulated more OptimusKG edges. The 80% condition MRR was 0.242 & 0.074, for PrimeKG FT and ZS, respectively. Additional Finetuning improved MRR.

Figure 4. Test & Validation Set Eccentricity

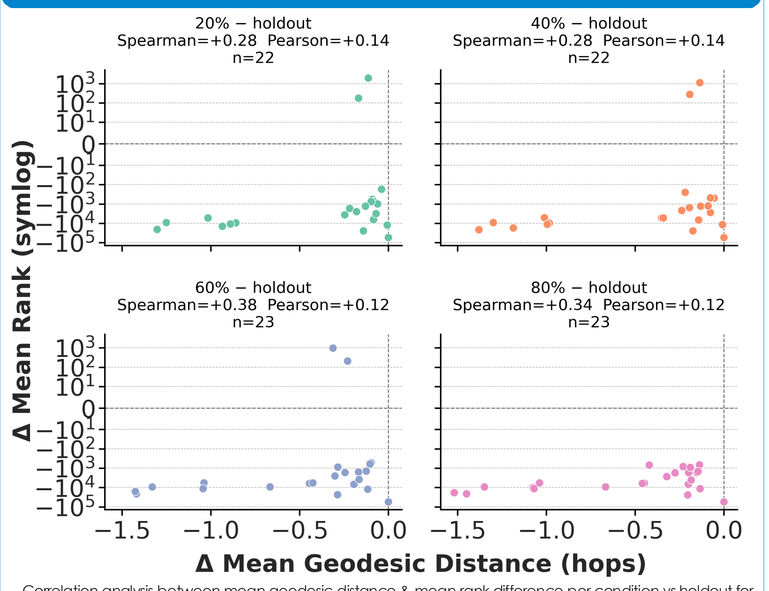


Upper: Test & Validation set eccentricity was measured for each condition. Eccentricity is the set of shortest paths between every point in the graph. Radius & diameter are the min & max eccentricity value. As more OptimusKG edges are added radius & diameter decrease. Lower: Eccentricity count plot for holdout & 80% conditions. Both are skewed right; 80% has better connectivity (shifted left).

Connectivity Differences between Conditions

- Geodesic distance (shortest path) & eccentricity between test/validation triples in each condition was evaluated to ascertain the effect connectivity has on model performance.
- As more OptimusKG triples are added to train, distance between concepts decrease in each condition (**Figure 4, upper**). Between the holdout & 80% condition, the radius (minimum) decreases from 12 to 10 & the diameter (maximum eccentricity) decreases from 23 to 18, respectively.
- The eccentricity distribution shifts leftward with a 2.47 hop mean difference & Cohen's *d* was 3.43 (paired) suggests a meaningful decrease in mean relative to the spread (σ) (**Figure 4, lower**).

Figure 5. Triple Type Distance & Rank Difference Correlation

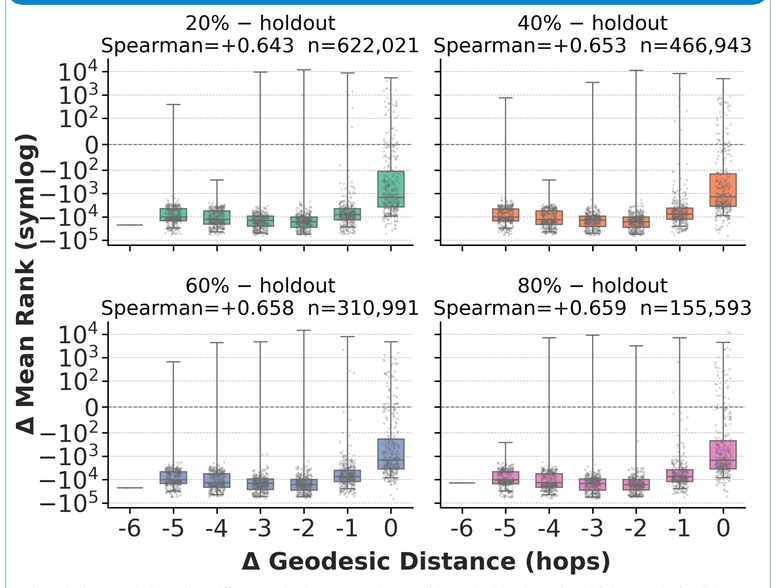


Correlation analysis between mean geodesic distance & mean rank difference per condition vs holdout for each triple type. Shortest paths for each triple was measured for each condition and holdout. The difference in shortest paths were correlated against the rank performance by the Finetuned – PrimeKG ULTRA checkpoint. Triples were aggregated by type.

Edge Accumulation Performance Analysis

- A correlation analysis between the change in each triple type specific mean distance & mean rank for the holdout & each condition (**Figure 5**). Across all conditions, a moderate monotonically positive (Spearman) & weak linear (Pearson) correlation was observed suggesting differences in connectivity among triple types influences model's predictive capacity.
- Correlation analysis on differences in geodesic distance & mean rank on unaggregated triples between each condition & holdout demonstrate moderately strong Spearman correlations. Each triple between each condition & holdout were paired; 50% of the test set was sampled for each condition (**Figure 6**). This suggests improvement in connectivity between head & tail entities potentially improve model's ranking performance.

Figure 6. Triple Distance & Rank Difference Correlation



Correlation analysis on the difference between each condition & holdout's paired triple geodesic distance & mean rank. A moderately strong Spearman correlation was observed cross all conditions; model's mean rank (Finetuned – PrimeKG checkpoint) improves as graph connectivity improves (decrease in hops).

References

1. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. Sci Data 10, 67 (2023).
2. Vittor, L., Noori A., Arango I., Polonuer J., Rodriques S., White A., Clifton D.A., & Zitnkik M. Unifying biomedical knowledge in a modern multimodal graph. Preprint at https://arxiv.org/abs/2604.27269
3. Galkin, M., Yuan, X., Mostafa, H., Tang, J. & Zhu, Z. Towards Foundation Models for Knowledge Graph Reasoning. Preprint at https://doi.org/10.48550/arXiv.2310.04562 (2024).
4. Aamer, N., Asim, M. N., Bhatti, A. I. & Dengel, A. Bind: large-scale biological interaction network discovery through knowledge graph-driven machine learning. J Transl Med 23, 856 (2025).

Acknowledgements

- The authors would like to thank Dr. Marcin von Grotthuss, Dr. Xinghao Zhang, Dr. Ali Ebrahimi, Dr. Maria Paraskevopoulou, Dr. Sun Rui, Dr. Jonghun Lee, Dr. Shih-Ying Wu, & Professor Marinka Zitnik, for their feedback. Special thanks to Dr. Yves Fomekong-Nanfack & Julie Bryant for their support.
- Disclosure Statement: This work was supported by Takeda Pharmaceutical Company Limited. The authors are employees of Takeda Pharmaceutical Company Limited and own company stock.

Scan the QR code to access environmentally friendly e-poster

Scan the QR code to access the poster and any additional materials on your mobile device.

No personal information is stored – standard charges may apply