Biologically Informed Graph Neural Network Construction via Pathway-Based Gene Interaction Graphs

Yuwei Xue, Sakib Mostafa, Lei Xing, Md Tauhidul Islam

MEDICAL PHYSICS COMMUNITY INSPIRATION INNOVATION IMPACT
2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

ABSTRACT

Biological systems are governed by structured molecular interactions, where regulatory circuits shape cellular behavior and drive disease progression. Much of this knowledge is naturally represented as graphs. However, most biomedical AI models cannot directly use graph-encoded biological knowledge and instead require compressed low-dimensional representations, which can lose important structure and reduce performance, especially in limited-sample clinical studies.

Here, we introduce Graph-in-Graph (GiG), a knowledge graph-modulated deep learning framework for data-efficient clinical prediction. GiG represents each patient as a standalone modular graph, in which curated biological knowledge graphs define edges and patient-specific measurements, such as gene expression, define node features. This design preserves gene-gene interactions and pathway topology in the model's learning process.

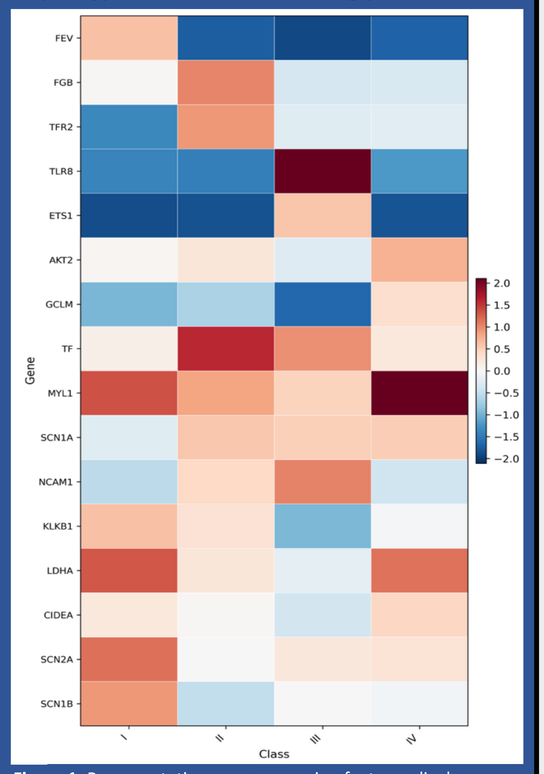


Figure 1. Representative gene-expression features display separable patterns across clinical groups in pan-cancer dataset.

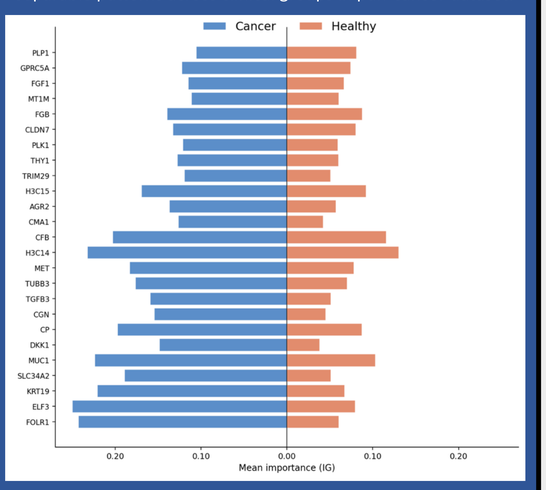


Figure 2. Top molecular features in cancer and healthy samples.

CONTACT

Md Tauhidul Islam
Stanford Medicine, Radiation Oncology
875 Blake Wilbur Drive, Stanford, CA
tauhid@Stanford.edu

INTRODUCTION

Biological systems are organized across multiple functional levels, with molecular interactions forming gene pathways that collectively determine cellular behavior and disease phenotypes. In these systems, genes participate in interconnected pathways, signaling cascades, structured regulatory circuits, and metabolic routes that have been mapped and recorded through decades of research in molecular and systems biology.

Modern genomic assays such as bulk transcriptomic, single-cell sequencing, and liquid biopsy now measure this molecular activity at high resolution across large patient cohorts, creating opportunities to connect gene-level measurements to clinical outcomes at a scale that was previously inaccessible.

However, most genomic machine learning pipelines today treat expression data as a flat table and ignore two foundational properties of the underlying biology. First, genes participate in curated interaction networks whose topology has been established through extensive experimental validation and carries real biological meaning. Second, the activity of a given pathway varies substantially between patients, so a single fixed network representation cannot capture how molecular signals are distributed across individuals within a disease cohort.

METHODS AND MATERIALS

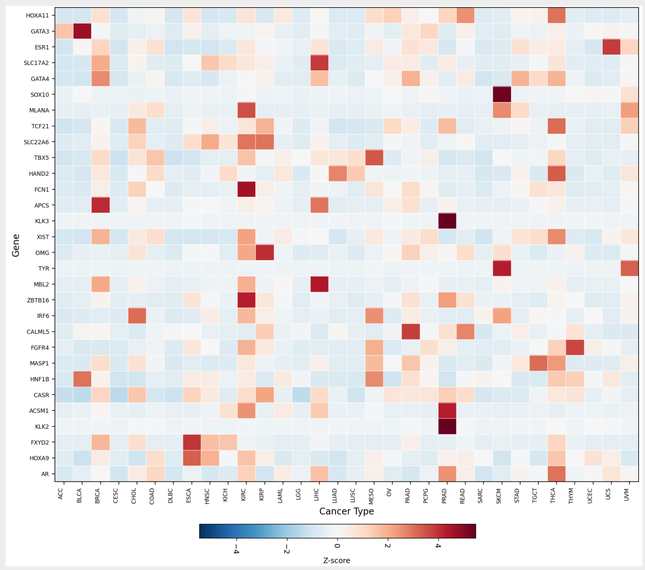


Figure 3. Heatmap of z-score normalized gene-level signals across 32 cancer types in the pan-cancer cohort.

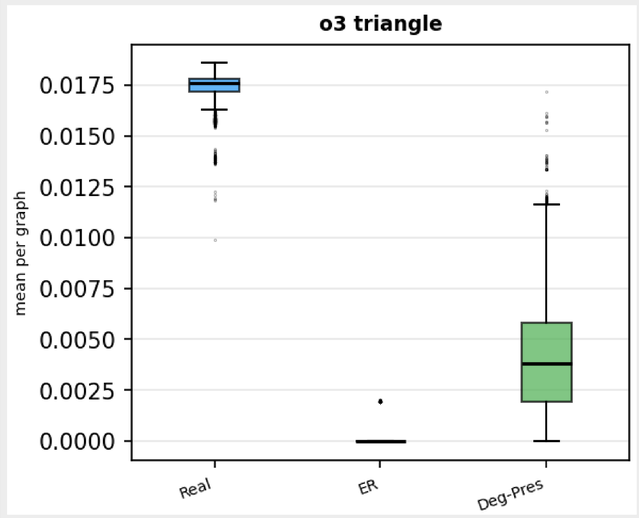
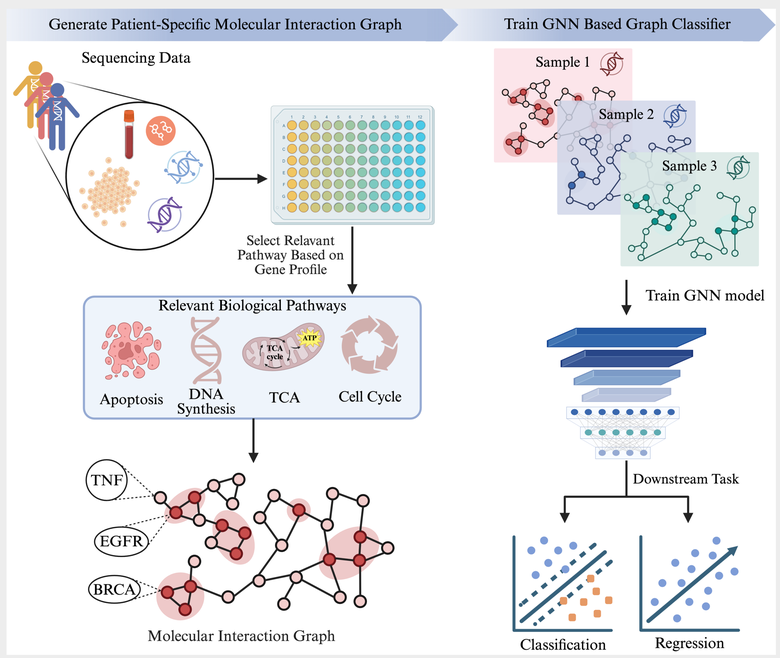


Figure 4. GiG-generated pathway graphs preserve distinct biological topology that contributes predictive signal beyond graph modeling alone.

Figure 5. GiG builds patient-specific pathway interaction graphs from omics data.



RESULTS

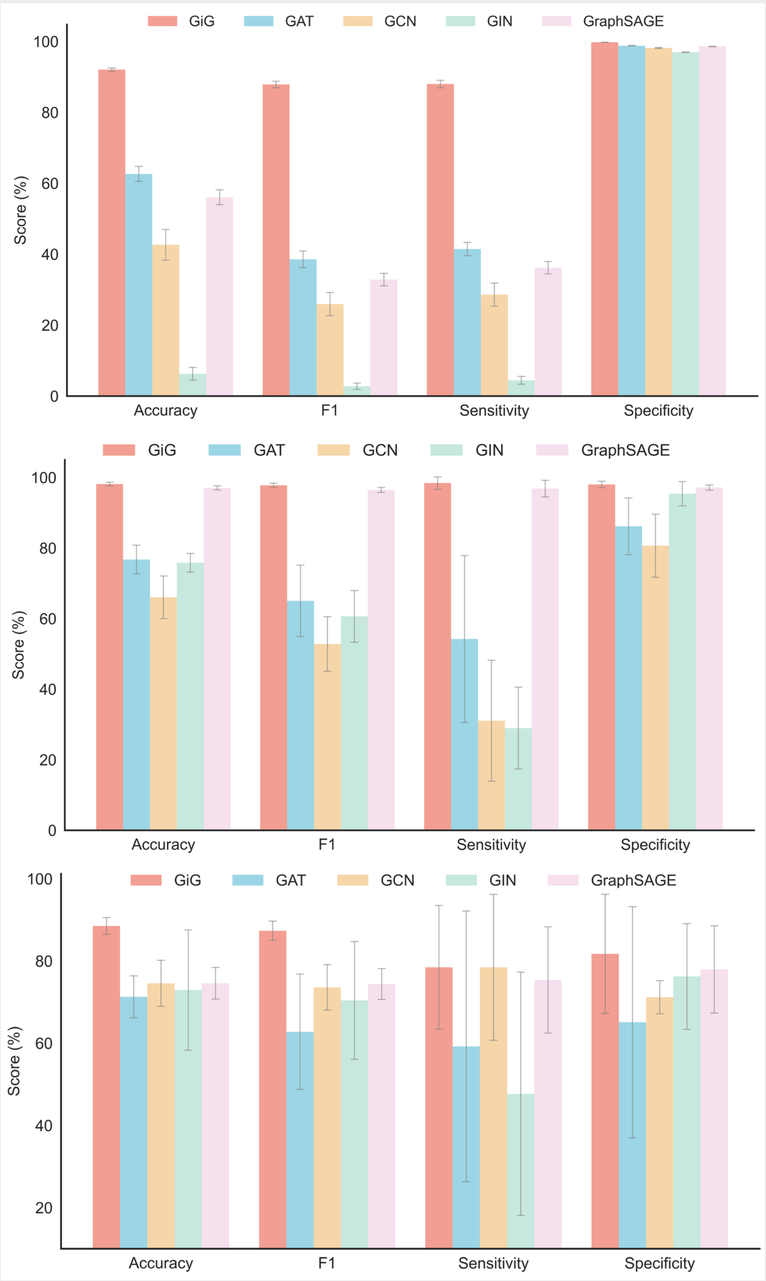


Figure 6. GiG outperforms baseline GNNs across pan-cancer classification, prostate cancer diagnosis, and RARE-Seq cfRNA cancer detection.

DISCUSSION

These results support the central hypothesis that organizing patient-specific transcriptomic measurements according to curated biological pathway topology can improve clinical prediction. The strongest benefit of pathway topology was observed in the most challenging prediction settings: low-signal cfRNA cancer detection and heterogeneous 32-class pan-cancer classification. In these settings, discriminative signals are likely distributed across multiple genes and pathway modules rather than concentrated in a few dominant markers. By aggregating information through biologically meaningful neighborhoods, GiG appears to improve robustness under noise, heterogeneity, and class imbalance.

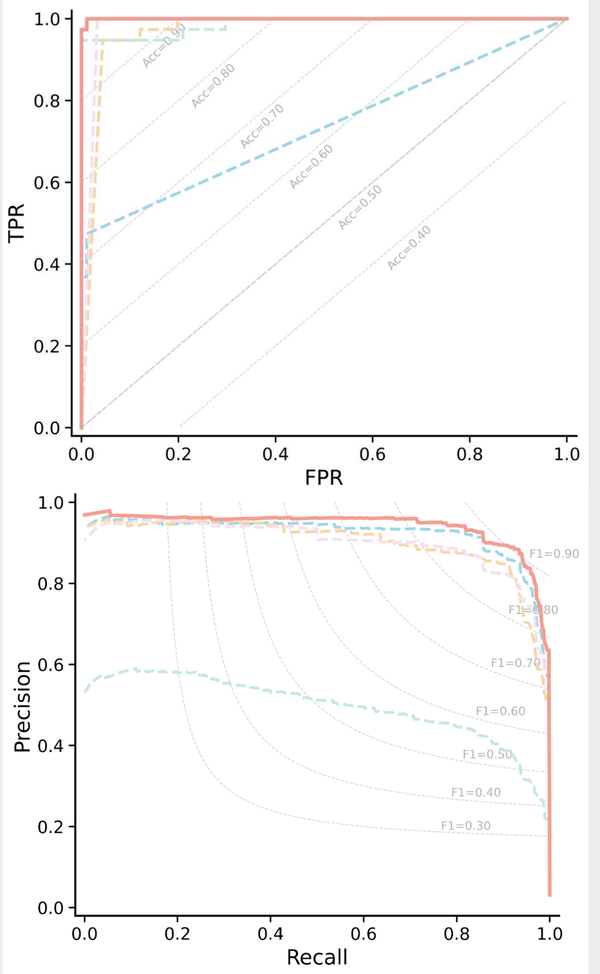


Figure 7-8. ROC-PR analysis shows stronger discrimination by GiG compared with conventional GNN baselines.

CONCLUSIONS

Across cohorts comprising nearly 9,700 patients and five clinical tasks, including liquid biopsy cancer detection, prostate cancer diagnosis, and 32-class pan-cancer classification, GiG consistently outperforms traditional and state-of-the-art methods, with the largest gains in limited-sample settings. Control experiments replacing real pathway graphs with random topologies confirm that these gains arise from biologically grounded knowledge graph structure rather than graph modeling alone. These findings show that knowledge graph-modulated deep learning can improve robustness, interpretability, and sample efficiency in clinical data analysis, and provide a principled framework for integrating biological knowledge graphs into predictive modeling.