Knowledge Graph Modulated Deep Learning for Limited-Sample Clinical Data Analysis

Yuwei Xue[1,], Sakib Mostafa[1,**], James Zou[2,3,4], Joseph Liao[5], Maximilian Diehn[1,6,7], Ash A. Alizadeh[8], Lei Xing[1,2,9], and Md Tauhidul Islam[1,*]**

[1]Department of Radiation Oncology, Stanford University, Stanford, California, USA
[2]Department of Electrical Engineering, Stanford University, Stanford, California, USA
[3]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, California, USA
[4]Department of Computer Science, Stanford University, Stanford, California, USA
[5]Department of Urology, Stanford University School of Medicine, Stanford, California, USA
[6]Stanford Cancer Institute, Stanford University, Stanford, California, USA
[7]Institute for Stem Cell Biology and Regenerative Medicine, Stanford University, Stanford, California, USA
[8]Department of Medicine, Division of Oncology, Stanford University, Stanford, California, USA
[9]Institute of Computational and Mathematical Engineering, Stanford University, Stanford, California, USA
[**]Equal contribution
[*]Correspondence to tauhid@stanford.edu

ABSTRACT

Biological systems are governed by structured molecular interactions, where pathways, regulatory circuits, and functional gene relationships shape cellular behavior and disease progression. Much of this critical biological knowledge is naturally represented as graphs. However, most biomedical AI models cannot directly integrate knowledge in graph form and instead require low-dimensional representations of biological knowledge. This compression often leads to information loss and substantially reduces model performance, especially in limited-sample settings frequently encountered in clinical studies. Here, we introduce Graph-in-Graph (GiG), a knowledge graph–modulated deep learning framework for data-efficient clinical prediction from limited patient samples. GiG represents each patient as a standalone modular graph, in which curated biological knowledge graphs can be integrated as edges and patient-specific information such as gene expression profiles can be integrated as node features. This flexible design enables the integration of multiple biological knowledge graphs while allowing the model to learn patient-level disease representations that preserve biologically meaningful information, including gene–gene interactions and pathway topology. Across several patient cohorts comprising nearly 11,000 patients and nine clinical tasks, including cancer detection from liquid biopsy dataset from Stanford Hospital, prostate cancer diagnosis, and a 32-class pan-cancer classification, GiG consistently outperforms traditional methods by a large margin, particularly in limited-sample settings. On the challenging prostate cancer diagnosis task, GiG improves macro-F1 performance by up to 49 percentage points when compared to state-of-the-art methods. Control experiments replacing real pathway graphs with random topologies confirm that the performance 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, providing a principled framework for integrating biological knowledge graphs into predictive modeling.

1 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 do not operate in isolation but 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[1,2]. 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[3,4]. Whether those opportunities translate into clinically useful predictive models depends in large part on computational frameworks that incorporate the organization of biology, rather than treating transcriptomic data as an unstructured list of values to be passed through a generic model[5,6].

Most genomic machine learning pipelines today continue to treat expression data as a flat table[7,8]. Gene values become columns in a matrix, and classifiers search for combinations of these columns that separate disease states from healthy ones.

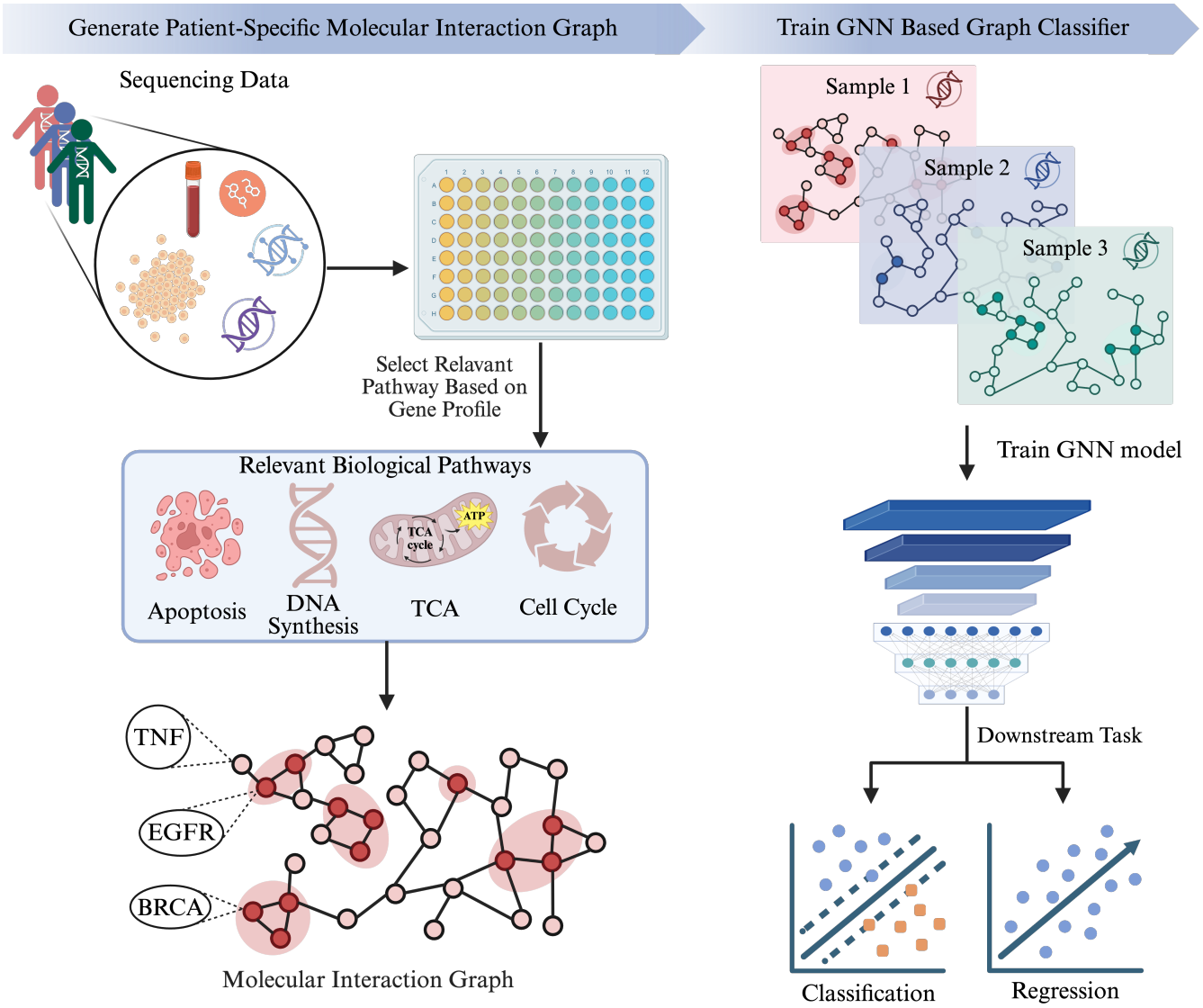

Figure 1. Workflow of the proposed GiG framework. GiG constructs patient-specific molecular interaction graphs by integrating transcriptomic profiles with curated biological pathway knowledge. Starting from sequencing-derived gene expression data, pathways relevant to each patient are selected based on the sample-specific gene expression profile. Curated pathway interactions are then merged to generate a molecular interaction graph in which nodes represent genes and edges encode known biological relationships. These patient-specific graphs are subsequently used as inputs to a graph neural network (GNN) classifier, enabling graph-level learning across a cohort of samples. The learned graph representations can then be applied to downstream predictive tasks such as disease classification and regression.

This framing provides convenient inputs for standard deep learning architectures, yet it ignores 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[9–11]. Second, the activity of a given pathway varies substantially between patients[2], so a single fixed network representation cannot capture how molecular signals are distributed across individuals within a disease cohort. Graph neural network methods developed for biomedical applications partially address this limitation by constructing patient-similarity networks from molecular features[12,13]. However, in these approaches, each patient is represented as a node with an associated feature vector rather than as an individual molecular interaction graph. Neither design allows pathway topology and patient-specific expression to interact at the level of message passing.

Several bodies of work have approached parts of this problem from different angles. Pathway-informed neural architectures constrain layer connectivity to follow curated biological groupings, producing models whose internal structure reflects canonical pathway annotations[14–17]. Graph neural networks applied to protein-protein interaction networks have shown that message passing over biological topology can improve performance on gene-level prediction tasks[18,19]. Patient-similarity networks combine clinical and molecular features to propagate label information across cohorts, leveraging the observation that phenotypically similar patients often share outcomes[20–22]. More recent work has extended graph foundation modeling to biomedical networks, learning transferable structural representations that can be adapted to multiple graph analysis tasks with limited labeled data[7,23,24]. However, currently there are no methods, in which every patient can be represented as a pathway-structured graph of their own, with the interaction topology integrated into each sample rather than supplied as an external prior shared across the cohort.

Here to address the limitation of existing methods and integrate biological knowledges as stand-alone graphs into deep learning pipelines, we introduce Graph-in-Graph (GiG), a framework in which each patient is represented as a pathway-structured graph and downstream tasks are performed across a population of such graphs. Nodes correspond to genes drawn from the patient transcriptomic profiles, and edges encode the topology of curated biological pathways sourced from WikiPathways[25,26], so that the functional organization of molecular interactions is encoded directly into the graph of every sample. Node features combine the patient gene expression values, a co-expression signal derived from sample-wise correlation structure, and a learnable gene-identity embedding that anchors each node to its molecular role across the cohort. Edge weights are set by the same correlation structure, allowing pathway topology to shape how information propagates during message passing while patient-specific expression controls what information actually travels along the edges. In this formulation, the lower-level pathway structure sits inside the higher-level patient graph, not as a separate preprocessing step but as an intrinsic component of the graph itself. Because the design is backbone agnostic, GiG can be instantiated with Graph Convolutional Network (GCN)[27], Graph Sample and Aggregate (GraphSAGE)[28], Graph Isomorphism Network (GIN)[29], Graph Attention Networks (GAT)[30,31], or any compatible graph neural network, which allows us to separate the contribution of the pathway-structured representation from the choice of message-passing architecture.

We evaluated GiG on several transcriptomic cohorts totaling close to 11,000 patients across nine classification tasks, including a liquid biopsy dataset (RARE-Seq)[32], a prostate cancer cohort[14], and a 32-class pan-cancer transcriptomic benchmark drawn from TCGA[33,34]. The most striking gain came from the pan-cancer classification task, which is the most challenging of the seven cohorts because of the limited patient samples per class. Across 32 cancer types, GiG achieved 92% accuracy and 88% macro F1, while the strongest conventional graph neural network baseline operating on the same inputs reached only 63% accuracy and 39% macro F1, corresponding to a gain up to 49% in macro F1 without any change to the underlying expression data or training protocol. The same pattern of consistent improvements over standard backbones held across the liquid biopsy and prostate cancer cohorts. To confirm that these gains come from the pathway topology itself rather than from node features alone, we replaced the real pathway graphs with Erdős Rényi and degree-preserving random controls and measured the resulting drop in classification accuracy[35,36]. We also characterized the structural distinctiveness of the real graphs using graphlet orbit statistics[37,38], which showed that pathway-derived topologies occupy a region of structural feature space that is clearly separated from both random-graph families across every dataset and task.

Beyond predictive performance, GiG's per-patient graph formulation provides a direct route to clinically meaningful interpretations. Because each patient is represented as an individual pathway-structured graph, each prediction can be traced back to the genes, pathways, and local graph structures that contribute most strongly to the output label. Such patient-specific molecular interpretations may support the generation of testable hypotheses concerning candidate biomarkers and disease-associated pathways[39]. Across patients with the same disease state, recurrent high-contribution pathways may identify shared molecular patterns for subsequent biological investigation[40–43].

In summary, here, we propose a per-patient graph formulation that integrates pathway topology directly into every sample, moving beyond the shared global network assumption that dominates current graph-based genomic modeling. With rigorous validation, we demonstrate consistent performance gains across seven heterogeneous cohorts and five classification tasks that differ in sample size, disease granularity, class cardinality, and class balance, showing that the benefits of pathway-structured modeling are not confined to a single benchmark or task type. We also validate through controlled structural perturbations and

graphlet orbit analysis that the observed gains arise from the pathway topology itself, not from the expression features in isolation. Because GiG's graph construction process is explicit and traceable, the framework is transparent at the patient, pathway, and gene levels. This makes it useful not only for classification but also for generating interpretable molecular hypotheses, including candidate biomarkers, pathway-level disease programs, therapeutic vulnerabilities, and potential mechanisms of treatment resistance[44]. Together, these results position GiG as a practical foundation for predictive modeling of transcriptomic data in precision medicine, in which the organization of biology is treated not as an afterthought to a flat feature vector but as a structural component of the model.

2 Results

2.1 GiG outperforms state-of-the-art GNNs in cancer subtype and stage classification on an ultrasensitive cfRNA dataset

We first evaluated GiG's ability to accurately detect cancer and predict cancer subtype and stage, of patient samples in the RARE-Seq cfRNA dataset from Stanford hospital[32]. RARE-Seq is designed to detect extremely low-abundance tumor-derived transcripts in plasma. Unlike tissue-based transcriptomic datasets, where tumor signal is relatively strong and localized, cfRNA measurements reflect a highly diluted and fragmented representation of tumor biology embedded within a background of circulating RNA from multiple tissue sources[45–47]. In other words, the RARE-Seq data set is characterized by low signal-to-noise ratios, substantial inter-patient heterogeneity, and significant class imbalance. Therefore, accurate prediction of cancer presence, subtype, and stage within this dataset requires a model to extract weak, distributed signals that are often obscured by noise and variability[45,47]. In this study, following traditional gene expression data pre-processing methods, we selected 10,000 highly variable genes (HVGs) out of over 60,000 genes from the RARE-Seq dataset to train and test our model. By focusing on genes that exhibit the greatest variability across samples, this setup allows us to capture biologically meaningful differences between disease states while reducing the influence of low-variance, noise-dominated features.

As shown in Figure 2e, GiG consistently outperforms node-level graph neural network baselines - including GAT, GCN, GIN, and GraphSAGE - across binary, stage, and subtype classification tasks (Fig. 2e, S2). Its advantage is most pronounced in the binary classification task, where GiG achieves the highest scores across all evaluation metrics. Specifically, GiG attains an accuracy of 89% and a macro-F1 score of 88%, outperforming the strongest baseline by over 15% in accuracy and 14% in macro-F1. GiG also demonstrates superior sensitivity (88%) and specificity (89%), indicating strong and balanced performance across both classes. In contrast, even the strongest baseline model, GCN, achieves only approximately 74% accuracy and macro-F1, with sensitivity and specificity around 78%. Overall, this indicates GiG has a more balanced ability to capture minority class signals under significant class imbalance, while node-based models are more prone to favoring dominant classes, a common challenge in cfRNA-based classification tasks[48,49]. As such, incorporating pathway-structured, sample-specific graph representations in our model enhanced its ability to retrieve weak but biologically meaningful signals embedded within noisy and heterogeneous cfRNA measurements. Together, these results demonstrate that GiG not only improves overall predictive accuracy but also achieves more robust and balanced performance across evaluation metrics, which is particularly critical in clinical settings where both false positives and false negatives carry significant consequences[50,51].

Beyond classification metrics, we evaluated GiG's performance using macro-averaged ROC curves. As shown in Figure 2c, GiG consistently outperforms node-level baselines across operating thresholds, with the largest separation observed at low false positive rates - a region of particular importance for high-precision classification tasks that are clinically relevant. In this crucial region, GiG achieves higher true positive rates than all baselines, indicating improved sensitivity under stringent false positive constraints. In contrast, node-level models have more gradual increases in the true positive rate and require substantially higher false positive rates to achieve comparable sensitivity.

To characterize the distribution of predictive signals across genes, we analyzed class-wise expression patterns for stage and subtype (Fig. 2a, d). Z-scored profiles reveal that each class exhibits distinct expression patterns across subsets of genes. For example, in Figure 2a, Stage I samples show elevated expression of SCN1B and reduced expression of TLR8, whereas Stage III demonstrates the opposite pattern, with lower SCN1B expression and higher TLR8 levels. These contrasting patterns indicate that disease stage identity is defined by distinct multi-gene expression signatures rather than the reliance on a single dominant marker. Similarly, Figure 2d shows distinct expression patterns across cancer subtypes, with each subtype characterized by a unique combination of up- and down-regulated genes. Notably, these differences do not reflect uniform shifts across all features but instead arise from heterogeneous changes across gene subsets. This suggests that discriminative signal in cfRNA data is distributed in a class-dependent manner, with distinct transcriptional patterns underlying different cancer subtypes.

Lastly, we examined Integrated Gradients (IG) attribution scores[52] using a butterfly plot to assess how predictive signals are allocated across classes in the binary task (Fig. 2b). The results reveal clear class-specific asymmetry in feature importance, with distinct sets of genes contributing preferentially to cancer and healthy predictions. Genes such as FOLR1[53], ELF3[54], and KRT19[55] exhibit elevated attribution for cancer classification but relatively low importance for healthy classification, while a different subset of genes shows higher attribution for healthy samples. The elevated cancer-specific attribution of FOLR1,

ELF3, and KRT19 is biologically consistent with their established roles as epithelial and tumor-associated markers. FOLR1 has been linked to tumor and circulating protein levels in ovarian cancer[56,57], ELF3 regulates epithelial transcriptional programs involved in tumor progression[54], and KRT19 is a canonical epithelial cytokeratin frequently used in circulating tumor cell detection[58]. These associations suggest that GiG prioritizes biologically plausible tumor-derived cfRNA signals for cancer classification rather than relying solely on nonspecific plasma RNA variation.

Taken together, these results demonstrate that incorporating pathway-structured representations allowed GiG to consistently improve predictive performance while capturing biologically meaningful and class-specific signal in a clinically challenging cfRNA setting. The observed gains in accuracy, sensitivity, and macro-F1, along with improved discrimination across ROC thresholds, indicate that pathway-structured representations enhance robustness under conditions of low signal-to-noise ratio, substantial heterogeneity, and class imbalance. The key advantage of GiG is not simply higher accuracy, but its ability to organize noisy circulating molecular measurements into patient-specific pathway graphs, allowing disease-relevant signals to emerge through biologically constrained message passing. Unlike pipelines that depend more heavily on assay-specific contamination removal, enrichment-score filtering, or manually selected disease markers, GiG leverages curated pathway topology to extract meaningful signals in a patient-specific manner[32]. In other words, rather than relying on extensive preprocessing to remove cfRNA noise, GiG demonstrates an ability to prioritize coherent tumor-associated molecular patterns over less informative variations in each patient. Supporting this interpretation, expression and attribution analyses show that GiG captures class-specific signals distributed across coordinated gene subsets, producing interpretable gene-level explanations that align with known cancer biology. Together, these findings demonstrate the value of pathway-structured graph learning for extracting clinically meaningful signals from noisy and heterogeneous cfRNA measurements, with potential applications in minimally invasive cancer detection and stratification.

2.2 GiG demonstrates near-perfect performance and robust generalization on prostate cancer diagnosis in a tissue-based transcriptomic dataset

Next, we evaluated GiG on a tumor tissue-derived prostate cancer transcriptomic data set[14], where tissue-based expression profiles capture more direct molecular programs associated with tumor development and progression. This dataset is characterized by well-defined molecular profiles and recurrent alterations in key signaling pathways that collectively shape prostate cancer biology, like PI3K/AKT, MAPK, and Wnt signaling. Therefore, this setting allows us to assess whether the advantages of pathway-structured representations become even more pronounced in a higher-signal regime, where biological organization is more clearly represented in the data, compared to the more diffuse and noisy signals observed in cfRNA.[14]

As shown in Figure 3e, GiG achieves near-perfect performance across all evaluation metrics, substantially outperforming node-level graph neural network baselines. GiG attains accuracy, macro-F1, and sensitivity values approaching 98%, with consistently low variance across folds. In contrast, node-level baselines - including GAT, GCN, and GIN - exhibit significantly lower and more variable performance, particularly in sensitivity, where several models fall below 60%. While GraphSAGE achieves competitive accuracy and specificity, GiG still performs better across all four metrics measured. These results indicate that GiG not only improves predictive accuracy but also maintains consistently high recall of the positive class, which is critical for identifying clinically relevant disease states.

Confusion matrix analysis further supports these findings, showing that GiG achieves near-perfect classification with minimal misclassification between localized and metastatic samples (Fig. 3b). Evidently, GiG is able to effectively capture discriminative patterns associated with disease progression. Consistent with this observation, macro-averaged ROC curves also demonstrate that GiG outperforms baseline models, achieving near-optimal true positive rates at extremely low false positive rates across operating thresholds (Fig. 3d). This indicates strong separability in the learned representations and highlights GiG's ability to operate effectively under stringent, clinically relevant classification criteria.

Lastly, to investigate the molecular features driving these predictions, we analyzed its Integrated Gradients (IG) attribution scores (Fig. 3a). The results reveal that predictive importance is distributed across a set of biologically relevant genes, including AKT1[59], MAPK1[60], and CTNNB1[61]. These genes are components of the PI3K/AKT, MAPK, and Wnt signaling pathways, which are frequently altered during prostate cancer development and progression[14,62,63]. Such alignment with established prostate cancer biology suggests that GiG is capable of prioritizing features consistent with underlying disease mechanisms.

GiG's ability to distinguish localized or primary prostate cancer from metastatic disease has clinical significance as localized and metastatic prostate cancer differ substantially in prognosis, treatment goals, and clinical management[64,65]. GiG's near-ceiling performance in disease-state classification highlights its potential to support existing prostate cancer diagnostic and clinical decision-making workflows. Furthermore, because the model's predictions can be traced to pathway- and gene-level contributors, GiG may also help identify biological pathways associated with disease progression, treatment resistance, or therapeutic vulnerability, positioning it not only as a predictive framework but also as a potential tool for risk stratification[66], biomarker discovery[67], and personalized treatment planning[68].

In summary, these results show that GiG prioritizes biologically meaningful genes associated with canonical prostate

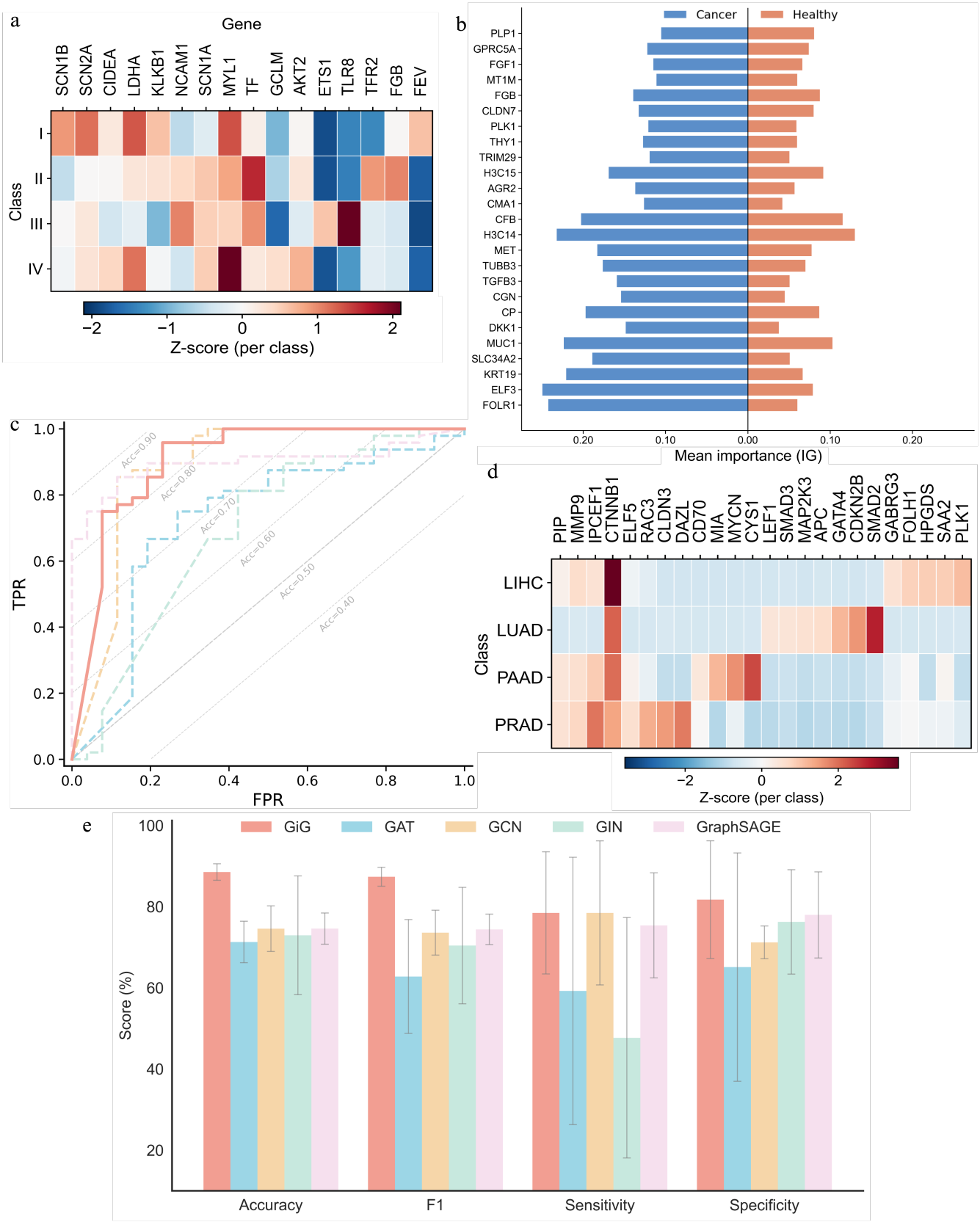

Figure 2. GiG improves cancer detection, prediction of cancer stages and subtypes and identifies class-dependent cfRNA signals in the RARE-Seq cohort. a. Stage-wise z-scored expression heatmap for representative genes. **b.** Integrated Gradients (IG) attribution butterfly plot for cancer detection. **c.** Macro-averaged ROC curves comparing GiG with GAT, GCN, GIN, and GraphSAGE methods. **d.** Subtype-wise z-scored expression heatmap for representative genes. **e.** Cross-validation performance across accuracy, macro-F1, sensitivity, and specificity. Bars indicate mean scores and error bars indicate variability across folds.

cancer pathways and achieves near-ceiling performance across accuracy, sensitivity, macro-F1, and ROC-based discrimination. Consistent with observations from the RARE-Seq dataset, pathway incorporation in GiG enables highly separable and reliable classification. In other words, GiG is not only robust in noisy, low-signal environments but achieves even stronger performance in higher-signal regimes, where biological organization is more clearly represented and can be more effectively leveraged for clinically useful applications.

2.3 GiG achieves scalable multi-class cancer classification across diverse pan-cancer types

Next, we evaluated GiG on a pan-cancer transcriptomic dataset derived from The Cancer Genome Atlas (TCGA), representing a large-scale, multi-class classification setting with substantial molecular and tissue-level heterogeneity[33]. Unlike the binary and stage-specific tasks considered previously, this dataset requires the model to distinguish a diverse set of cancer types, each characterized by different transcriptional programs with varying degrees of similarity. As such, this dataset provides a stringent test of whether GiG can scale effectively and maintain performance in the presence of increased class complexity and inter-class variability.

As shown in Figure 4e, GiG achieves over 90% accuracy and steadily outperforms node-level graph neural network baselines across all evaluation metrics. Consistent with observations from both the RARE-Seq and prostate cancer datasets, GiG maintains substantially higher macro-F1 and sensitivity scores than GAT, GCN, GIN, and GraphSAGE, which demonstrates its superior ability to maintain balanced predictions across diverse cancer classes. Although specificity remains relatively similar across models, GiG still achieves the strongest overall performance. Together, these results show that GiG is able to provide robust and accurate classification even in a large-scale, multi-class setting characterized by substantial biological heterogeneity.

Confusion matrix analysis further illustrates GiG's ability to distinguish between cancer types with high fidelity, showing strong diagonal dominance and minimal confusion across classes (Fig. 4b). This suggests that the model effectively captures discriminative features unique to each cancer type, despite the biological overlap between certain classes. Consistent with these observations, macro-averaged ROC curves demonstrate that GiG outperforms baseline models across operating thresholds, maintaining superior true positive rates at low false positive rates even in this multi-class setting (Fig. 4c). Similarly, precision-recall curves show that GiG achieves consistently higher precision across a wide range of recall values compared to baseline models (Fig. 4d), indicating improved performance under class imbalance, reinforcing its ability to make reliable predictions across diverse cancer types.

To further examine the molecular features driving pan-cancer classification, we analyzed Integrated Gradients (IG) attribution scores across cancer types using a heatmap representation (Fig. 4a). As shown in Figure 4a, each cancer type exhibits distinct attribution patterns across different subsets of genes. Evidently, GiG distributes predictive importance in a class-dependent manner, rather than relying on a shared set of dominant features across all cancer types. These results suggest that pathway-informed graph representations enable GiG to capture discriminative molecular signals specific to individual cancer types despite extensive transcriptomic heterogeneity. Several genes also showed cancer-type-specific attribution patterns. APCS[69] received elevated attribution in liver hepatocellular carcinoma, ZBTB16[70] and MLLT6[71] in kidney cancers, and TYR in uveal melanoma[72]. These findings demonstrate that GiG assigns different predictive contributions across cancer types, although the biological and clinical significance of these associations requires further validation.

Together, these results demonstrate that GiG effectively scales to complex multi-class classification tasks, maintaining strong performance, accurate and balanced predictions across diverse cancer types. Furthermore, GiG's ability to capture class-specific gene expression patterns whilst achieving superior accuracy and precision indicates that pathway-informed graph representations provide a robust framework for modeling large-scale, heterogeneous transcriptomic datasets. Demonstrably, the advantages of GiG extend beyond binary classification task shown in RARE-Seq and classification task in Prostate Cancer dataset: it can be applied to a large multiclass problem while producing gene-level attribution patterns that can be investigated in subsequent biological studies.

2.4 GiG achieves superior tumor stage classification in urothelial carcinoma

To further evaluate GiG across clinically relevant disease states in an independent tissue-based cohort, we applied the framework to GSE32894[73], a urothelial carcinoma transcriptomic cohort comprising 308 fresh-frozen tumor resection specimens with tumor stage annotations. Unlike binary cancer detection, tumor stage classification requires the model to distinguish biologically related disease states that represent differences in the extent and invasiveness of the same malignancy. In this analysis, samples were classified across Ta, T1, and T2 tumor stages. Urothelial carcinoma is characterized by substantial molecular and clinical heterogeneity[74–76], and this task provides an additional test of whether pathway-structured representations can capture graded molecular changes associated with disease progression.

As shown in Figure 5e, GiG achieves the strongest overall performance across accuracy, macro-F1, sensitivity, and specificity among the evaluated models. GiG reaches approximately 65% accuracy and F1, with sensitivity at a similar level and specificity above 80%. GraphSAGE performs comparably, whereas GAT, GCN, and GIN show lower performance across most metrics. Consistent with the performance metrics, the macro-averaged ROC curves show that GiG and GraphSAGE provide

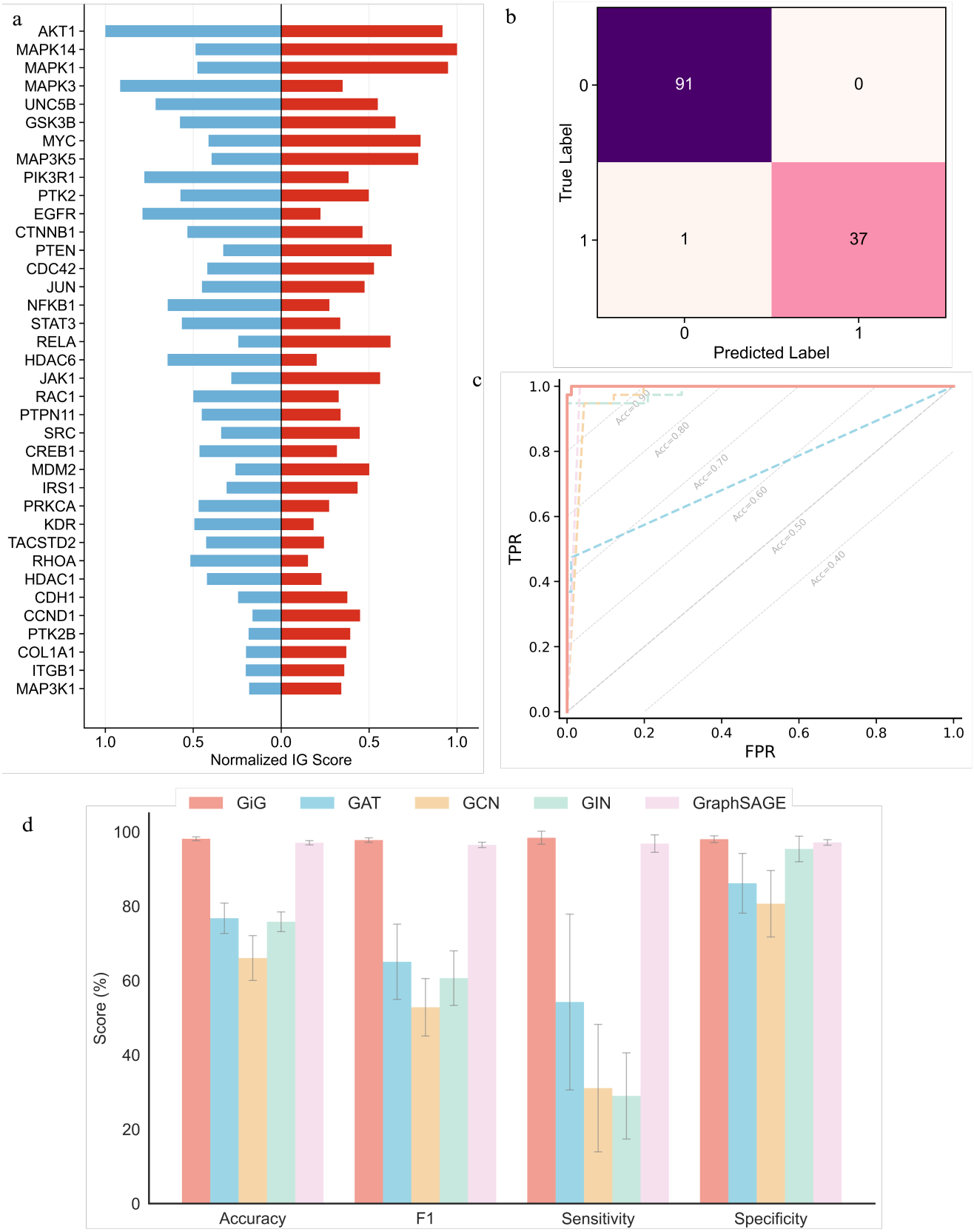

Figure 3. GiG achieves near-perfect performance in prostate cancer classification. a. Integrated Gradients (IG) attribution plot highlighting predictive genes, including canonical pathway components such as AKT1, MAPK1, and CTNNB1. **b.** Confusion matrix for localized versus metastatic prostate cancer classification. **c.** Macro-averaged ROC curves comparing GiG with GAT, GCN, GIN, and GraphSAGE baselines. **d.** Cross-validation performance across accuracy, macro-F1, sensitivity, and specificity. Bars indicate mean scores and error bars indicate variability across folds.

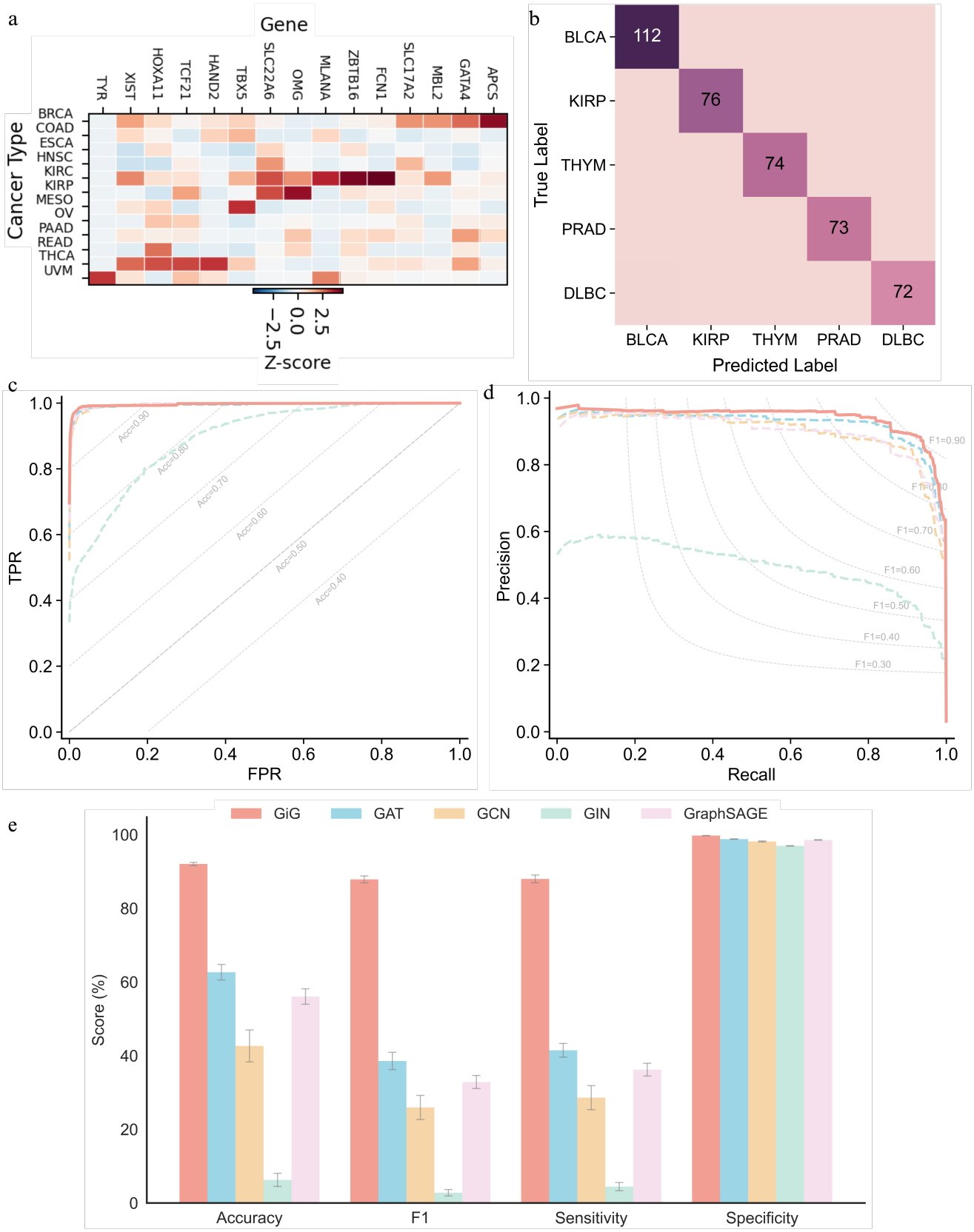

Figure 4. GiG achieves scalable multi-class classification across pan-cancer types. a. Z-scored Integrated Gradients attribution heatmap shows class-specific gene importance patterns across cancer types. **b.** Confusion matrix for the top five cancer classes demonstrates strong diagonal dominance and limited inter-class misclassification. **c.** Macro-averaged ROC curves comparing GiG with node-level GNN baselines. **d.** Macro-averaged precision-recall curves showing improved precision across recall thresholds. **e.** Cross-validation performance across accuracy, macro-F1, sensitivity, and specificity. Bars indicate mean scores and error bars indicate variability across folds.

the strongest discrimination across operating thresholds, with GiG showing an advantage particularly at lower false-positive rates (Fig. 5d). Precision-recall analysis similarly places GiG among the best-performing models across recall values (Fig. 5c). Overall, these results suggest that pathway-structured patient representations retain an advantage in distinguishing urothelial tumor stages, with varying magnitude of this advantage depending on the underlying dataset and prediction task.

The confusion matrix further illustrates the ability of GiG to distinguish among the three tumor stages (Fig. 5b). As shown along the diagonal, GiG correctly classifies 91 of 116 Ta tumors (78.4%), compared with 54 of 97 T1 tumors (55.7%) and 50 of 82 T2 tumors (61.0%) (Fig. 5b). The stronger classification of Ta tumors is consistent with the underlying biological differences across these three pathological stages: Ta tumors are predominantly associated with the Urobasal A molecular subtype, whereas T1 and T2 tumors are more invasive and exhibit much greater molecular heterogeneity across molecular subtypes[73]. Therefore, GiG's lower class-wise performance for T1 and T2 tumors is consistent with the challenge of inferring pathologically defined stage from transcriptomic measurements alone. In this context, GiG's superior overall performance shown in Figure 5e suggests that pathway structure may help recover discriminative signal even when stage-specific transcriptional boundaries are less distinct.

Next, to examine whether the stage predictions correspond to distinct transcriptional patterns, we compared class-wise expression profiles for representative genes (Fig. 5a). The resulting heatmap reveals stage-dependent expression patterns distributed across distinct sets of genes, with several genes exhibiting opposing expression profiles across the three classes. For example, PHGDH and TMEM59 show relatively elevated expression in T1 tumors and lower expression in T2 and Ta tumors, whereas BCYRN1 is preferentially elevated in T2 while CSRP1 shows a distinct Ta-associated pattern. Notably, PHGDH has previously been associated with high-grade and progressive bladder cancer, while BCYRN1 has been implicated in bladder cancer invasion and lymphatic metastasis[77,78]. Though these genes are not established stage-specific biomarkers, their non-monotonic expression patterns are consistent with the known molecular heterogeneity of urothelial carcinoma[73]. GiG's performance in this setting demonstrates its ability to integrate such distributed, stage-associated signals across pathologically defined tumor stages.

Taken together, the GSE32894 results extend GiG beyond the primary cohorts by demonstrating strong performance on a clinically relevant staging problem within urothelial carcinoma. Although the performance advantage is less conspicuous than that observed in the more heterogeneous pan-cancer and low-signal cfRNA tasks, these findings nonetheless highlight the potential of pathway-based aggregation to integrate transcriptional signals into coherent representations of disease states.

2.5 GiG demonstrates robust and consistent blood-based cancer detection from tumor-educated platelets

We evaluated GiG on GSE68086[79], a blood-based RNA-sequencing cohort derived from tumor-educated platelets (TEPs). The dataset contains platelet transcriptomic profiles from patients with multiple malignancies and healthy individuals, providing a complementary liquid biopsy setting to the RARE-Seq cfRNA cohort. Unlike circulating cell-free RNA, TEPs capture cancer-associated transcriptional changes indirectly through interactions between tumors and circulating platelets[80,81]. Therefore, successful classification requires the model to identify distributed cancer-associated molecular patterns within a blood cell population, rather than directly detecting transcripts released from tumor tissue itself. This provides an additional test of whether GiG can generalize across distinct liquid biopsy modalities and sources of circulating transcriptomic information.

As shown in Figure 6d, GiG achieves strong cancer-versus-healthy classification performance, attaining the highest accuracy among the evaluated models and remaining competitive with the strongest baselines across all metrics. GiG reaches over 90% mean accuracy and roughly 85% performance in F1, sensitivity, and specificity. Against the most competitive baseline model GraphSAGE, GiG exhibits substantially tighter error bars across evaluation folds. Therefore, the advantage observed in this cohort lies not only in GiG's high predictive performance, but also in its markedly greater consistency in maintaining such high performance across folds.

The confusion matrix further demonstrates balanced discrimination between cancer-associated and healthy platelet samples (Fig. 6b). GiG correctly identifies 210 of 224 cancer samples (93.8%) and 44 of 54 healthy samples (81.5%), showing that the model maintains substantial sensitivity to cancer-associated platelet signals without achieving its overall accuracy solely through prediction of the majority cancer class. Consistent with these findings, ROC analysis shows strong separation between cancer and healthy samples across operating thresholds. As shown in Figure 6c, GiG maintains high true positive rates at low false positive rates and remains among the strongest models across the ROC curve.

To investigate the molecular signals contributing to TEP-based cancer detection, we analyzed Integrated Gradients attribution scores for cancer and healthy classifications (Fig. 6a). The resulting butterfly plot reveals pronounced class-dependent asymmetry in predictive importance, with attribution distributed across a broad set of genes rather than concentrated in a single dominant feature. Notably, several highly attributed genes are consistent with biological processes previously implicated in cancer-associated platelet remodeling. In particular, SYK is a central mediator of platelet activation downstream of GPVI and CLEC-2 and participates heavily in tumor-platelet signaling[82,83], whereas TYROBP is associated with myeloid and inflammatory signaling[84]. This is consistent with recent evidence that TEP transcriptomes contain both activated-platelet and inflammatory-

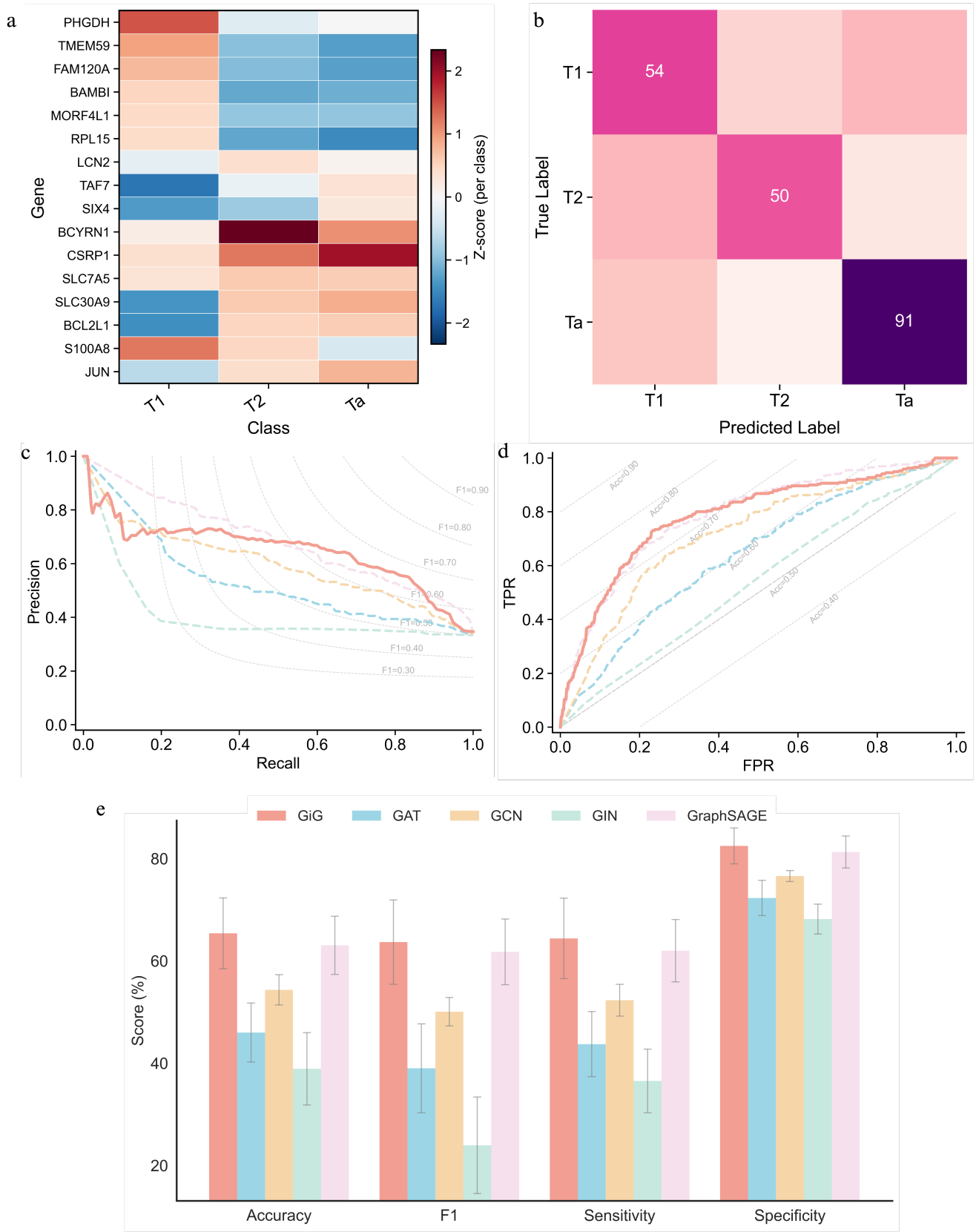

Figure 5. GiG improves tumor stage classification in urothelial carcinoma. a. Stage-wise z-scored expression heatmap for representative genes across Ta, T1, and T2 tumors in the GSE32894 cohort. **b.** Confusion matrix for multiclass tumor stage classification. **c.** Macro-averaged precision-recall curves comparing GiG with GAT, GCN, GIN, and GraphSAGE baselines. **d.** Macro-averaged ROC curves comparing model performance across tumor stages. **e.** Cross-validation performance across accuracy, macro-F1, sensitivity, and specificity. Bars indicate mean scores and error bars indicate variability across folds.

cell-associated signatures[85]. Other genes like DBI, ATM, and EPOR also contribute to the predictive representation, although their specific roles in tumor-educated platelets remain less well established[86,87]. Overall, the heterogeneity in these attributed genes is consistent with the diverse functionalities of TEP-associated transcriptional signals, spanning platelet activation, systemic inflammation, hematopoietic regulation, and broader cancer-associated molecular programs.

Taken together, our results demonstrate that GiG provides robust and consistent cancer detection from tumor-educated platelet RNA. While RARE-Seq measures sparse circulating cell-free RNA signals embedded within a heterogeneous extracellular RNA background, GSE68086 captures cancer-associated changes within the platelet transcriptome. Despite differences in assay modality and biological source, GiG maintains high and stable classification performance in both settings, supporting the broad applicability of patient-specific pathway graphs to liquid biopsy analysis, a clinically relevant and minimally invasive approach for cancer detection and monitoring[88].

3 Discussion

In this study, we introduced Graph-in-Graph (GiG), a patient-specific graph learning framework that integrates transcriptomic measurements with curated biological pathway topology. Across liquid biopsy, tissue-based prostate cancer, and pan-cancer classification tasks, GiG consistently improved predictive performance relative to standard graph neural network baselines. These results suggest that representing each patient as an individualized pathway-structured graph can improve clinical prediction from transcriptomic data, particularly when discriminative signals are distributed across multiple genes and pathway modules rather than concentrated in a small number of dominant markers.

A central question is whether pathway topology contributes structure beyond what node features already encode. We replaced the WikiPathways-derived edges with two random graph controls while keeping every other component of the model identical. Erdős–Rényi random graphs were generated while matching the number of nodes and edges in the corresponding pathway graph[89]. Degree-preserving rewiring retained the degree of each node while randomizing connectivity elsewhere[90]. The performance loss under each control measures how much the classifier depends on pathway topology for each task. Graphlet orbit statistics were computed for the pathway-derived graphs and both random controls using ORCA[91]. These analyses showed that the pathway-derived topology contained local structural patterns that were not reproduced by either random-graph family (Fig. S9, Fig. S6, Fig. S10).

On RARE-Seq binary cancer detection, the real pathway graph reached 89.2% accuracy and 87.7% macro-F1. Erdős–Rényi rewiring reduced performance to 73.0% accuracy and 72.5% macro-F1. Degree-preserving rewiring reduced performance further to 67.6% accuracy and 56.0% macro-F1. Disrupting pathway topology cost between 16.2% and 21.6% in accuracy and between 15.2% and 31.7% in macro-F1. Neither the edge count nor the degree sequence alone recovered the pathway signal. In the low-signal cfRNA setting, pathway structure organizes weak patient-specific molecular signals into functional neighborhoods where coordinated dysregulation aggregates more effectively than across random connectivity. The subtype and stage tasks reveal the same pattern through macro-F1 rather than accuracy. For subtype classification, the real graph reached 93.8% accuracy and 54.2% macro-F1, while both random controls held at 87.5% accuracy but collapsed to 23.3% macro-F1. For stage classification, the real graph reached 83.3% accuracy and 52.0% macro-F1, while both controls reached 77.1% accuracy and 21.8% macro-F1. The accuracy drop of 6.3% in both tasks understates the loss. The control models default to majority-class prediction, and the 87.5% subtype control accuracy matches the LUAD prevalence in the cohort to the nearest decimal. Macro-F1 collapses by 30.8% on subtype and 30.2% on stage. Pathway topology contributes most where the model must separate clinically meaningful subclasses rather than recover the dominant class.

The pan-cancer ablation extends this result across 32 cancer types. The real pathway graph reached 92.4% accuracy and 88.9% macro-F1. Erdős–Rényi rewiring reduced accuracy by 24.2% and macro-F1 by 31.8%. Degree-preserving rewiring reduced accuracy by 15.3% and macro-F1 by 18.7%. Random topologies retain nontrivial predictive ability, confirming that expression-derived node features carry disease-discriminative information on their own. Neither control approached the real-graph result. Across 32 cancer classes with substantial transcriptional overlap, curated pathway topology provides an inductive bias that the model cannot recover from node features alone. In contrast, the prostate cancer ablation showed only modest dependence on exact pathway topology. Relative to the real pathway graph, Erdős–Rényi random wiring reduced accuracy by only 1.55 percentage points and macro-F1 by 1.87 percentage points, while degree-preserving rewiring reduced accuracy by only 0.78 percentage points and macro-F1 by 0.94 percentage points. Such small decreases suggest that, in the prostate cancer cohort, transcriptomic node features already contained highly separable disease signal, making the model less sensitive to the precise arrangement of pathway edges. This result is useful because it prevents overclaiming: GiG does not require that pathway topology dominate in every dataset. Rather, the relative contribution of topology depends on the strength of the molecular signal and the difficulty of the classification task.

Together, these experiments separate the contribution of molecular signal from the contribution of biological structure. Across all datasets, random graphs retained nontrivial predictive ability, confirming that patient-specific expression features remain the primary source of discriminative information. However, the magnitude of performance loss after topology disruption

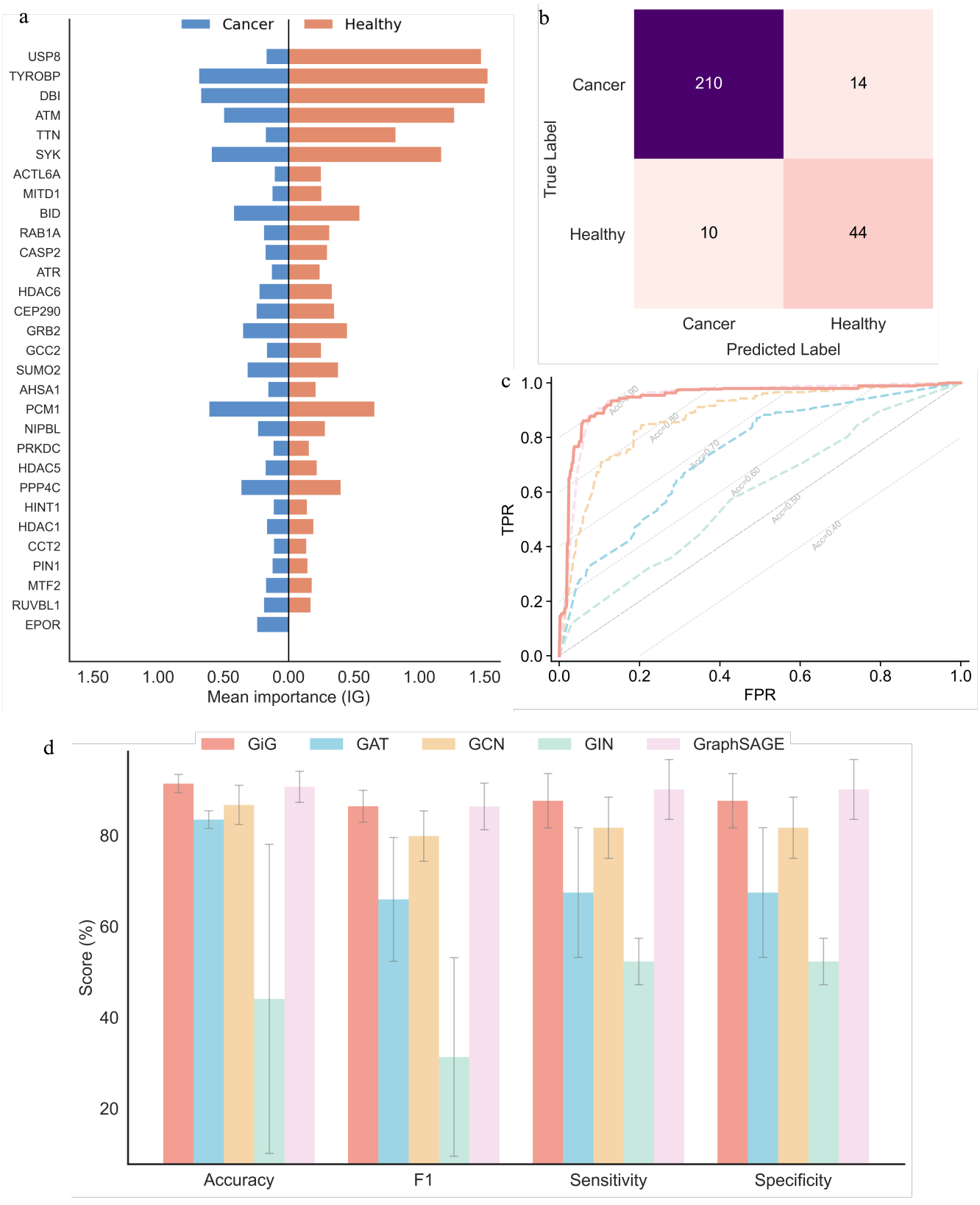

Figure 6. GiG enables accurate blood-based cancer detection from tumor-educated platelets. a. Integrated Gradients (IG) attribution butterfly plot highlighting class-dependent gene importance for cancer and healthy samples in the GSE68086 tumor-educated platelet cohort. **b.** Confusion matrix for binary classification of tumor-educated platelets from patients with cancer versus platelets from healthy individuals. **c.** ROC curves comparing GiG with GAT, GCN, GIN, and GraphSAGE baselines. **d.** Cross-validation performance across accuracy, F1 score, sensitivity, and specificity. Bars indicate mean scores and error bars indicate variability across folds.

shows that curated pathway structure contributes meaningful organization, especially in PanCancer and RARE-Seq, where biological heterogeneity and class imbalance make the learning more difficult. The strongest interpretation is therefore not that graph topology alone predicts cancer state, but that biologically meaningful topology improves how molecular measurements are represented, propagated, and aggregated.

GiG's performance improvements suggest several potential clinical applications. In liquid biopsy analysis, where tumor-derived cfRNA signals are sparse, diluted, and confounded by non-tumor background noise, pathway-structured graphs can help organize weak molecular signals into biologically coherent neighborhoods, reducing dependence on assay-specific feature engineering. In prostate cancer, GiG's ability to distinguish between localized and metastatic disease states shows clinical potential in improving prognoses, treatment choice, and disease monitoring. GiG is not intended to replace histopathology, imaging, prostate-specific antigen (PSA) kinetics, or other exising clinical procedures, but it has potential to be a decision-support tool downstream. In pan-cancer classification, GiG's strong performance across 32 cancer types suggests that patient-specific pathway graphs can preserve disease-defining molecular signatures across highly heterogeneous tumor contexts. Across all three settings, the results motivate further evaluation of GiG in independent and prospectively collected clinical cohorts. To further validate these findings, we also obtained data from the GEO database[92], including GSE32894 (n=308), GSE39582 (n=585), GSE62254 (n=300), and GSE68086 (n=228). Supplementary analyses across these four additional transcriptomic datasets further supported the generalizability of this framework across cancer types, assay modalities, and clinically relevant prediction tasks, including tumor progression, grade and stage, BRAF mutation status, tumor location, epithelial-mesenchymal phenotype, and blood-based cancer detection from tumor-educated platelets (Fig. S17 - Fig. S22). Together, these results suggest that GiG can capture pathway-level transcriptional patterns associated not only with overall cancer identity, but also with more nuanced and clinically meaningful differences in disease aggressiveness, molecular state, and progression.

Overall, the experiments support the central hypothesis that biologically grounded pathway topology contributes to GiG's performance, but they also show that its contribution varies with the structure of the prediction task. The largest gains from real pathway topology were observed in the settings where biological signal is most difficult to decode, including low-signal cfRNA classification and heterogeneous pan-cancer classification. These results suggest that patient-specific pathway graphs can serve as a useful inductive bias for clinical transcriptomic modeling, improving robustness and interpretability while preserving the molecular organization of disease biology. As larger and more diverse clinical omics datasets become available, GiG provides a flexible framework for integrating curated biological knowledge directly into patient-level predictive models.

4 Methods

Graph-in-Graph (GiG) was designed to convert transcriptomic measurements into patient-specific molecular interaction graphs that could be used directly for graph-level learning and predictions. The graph construction pipeline consisted of four main stages: gene identifier canonicalization, sample-specific selection of transcriptionally dysregulated genes, mapping of selected genes to curated biological pathways, and assembly of a patient-level graph by merging the corresponding pathway interaction graphs. The resulting graph for each patient encoded both sample-specific molecular activity and prior biological pathway structure. These graphs were then used as inputs to graph neural network classifiers for downstream disease classification tasks.

4.1 Transcriptomic preprocessing and gene identifier canonicalization

For each dataset, raw expression matrices were first reformatted so that rows corresponded to genes and columns corresponded to samples during graph construction. Gene identifiers were canonicalized to HUGO Gene Nomenclature Committee (HGNC) gene symbols[93] to ensure consistent alignment between the measured transcriptomic expression matrix and WikiPathways-derived pathway topology. For datasets with Ensembl[94] gene identifiers, version suffixes were removed before identifiers were queried using MyGeneInfo[95], and only genes annotated as protein-coding with valid HGNC[96] symbols were retained. Genes without valid HGNC mappings were excluded from downstream graph construction, ensuring that every graph node corresponded to a uniquely resolved, measured gene feature, reducing the risk of introducing unmatched or biologically ambiguous nodes into the patient-specific pathway graphs.

To mitigate the effect of low-variance and less informative features, highly variable genes were selected before graph construction. Specifically, genes were ranked by variance across samples, and the most highly variable genes were retained for subsequent pathway mapping. In the implemented graph-construction pipeline, the RARE-Seq and prostate cancer cohorts were filtered to the top 10,000 highly variable genes before sample-specific graph assembly, whereas the pan-cancer dataset used the pre-aligned gene universe provided as input. The pan-cancer cohort was not subjected to additional HVG selection because the input matrix was already provided as a pre-aligned gene expression matrix for cross-cancer comparison, and introducing an additional filtering step could remove subtype specific genes with lower global variance but potential relevance to individual cancer classes. After preprocessing, the resulting expression matrix was denoted as

$$X \in \mathbb{R}^{G \times N},$$

where G is the number of retained genes and N is the number of samples.

4.2 Sample-specific selection of dysregulated genes

For each sample, genes were selected on the basis of their relative expression deviation within that sample. Let x_{gi} denote the expression value of gene g in sample i. Expression values were standardized independently within each sample across all retained genes:

$$z_{gi} = \frac{x_{gi} - \mu_i}{\sigma_i},$$

where μ_i and σ_i are the mean and standard deviation of expression values across all genes in sample i. The absolute z-score,

$$m_{gi} = |z_{gi}|,$$

was used as a direction-independent measure of transcriptional dysregulation, allowing both highly up-regulated and highly down-regulated genes to contribute to graph construction.

For each sample i, a dysregulated gene set D_i was defined by retaining genes whose absolute z-score exceeded a dataset-specific percentile threshold:

$$D_i = \{g \in G : |z_{gi}| \geq Q_\tau(|z_i|)\},$$

where $Q_\tau(|z_i|)$ is the τ-th percentile of absolute z-score values across genes in sample i. For the primary GiG pipeline, patient-specific graphs were constructed using an 80th percentile threshold on the absolute z-score distribution. This threshold retained the most dysregulated genes in each sample while preserving sufficient pathway coverage for graph construction. Alternative percentile thresholds, including 10th, 20th, and 90th percentile cutoffs, were also evaluated during pipeline development: the 80th percentile threshold was selected for the final analyses because it provided a practical balance between sample-specific sparsity and biological pathway representation.

4.3 Mapping dysregulated genes to WikiPathways

Selected genes were mapped to curated biological pathways using WikiPathways[25,26]. For each gene $g \in D_i$, pathway membership was queried in WikiPathways using the human gene identifier namespace. The returned pathway identifiers were represented as WikiPathways IDs (WPIDs). For sample i, the set of selected pathways was defined as

$$P_i = \bigcup_{g \in D_i} P(g),$$

where $P(g)$ denotes the set of curated pathways containing gene g in the WikiPathways database. Genes without associated pathways were recorded and excluded from pathway selection. Gene-to-pathway mappings were cached locally to avoid repeated queries and to ensure that identical gene inputs yielded the same pathway set across experiments. Overall, this procedure allowed each sample to select a distinct set of pathways based on its unique transcriptional profile. As a result, the final graph topology was not fixed across all samples, but instead reflected the pathways associated with the most dysregulated genes in each individual sample.

4.4 Construction of pathway-level interaction graphs

For each WikiPathways Identifier (WPID) selected for a sample, the corresponding pathway was retrieved from WikiPathways in Graphical Pathway Markup Language (GPML) format and cached locally for reuse. Each GPML file was parsed to extract two classes of pathway elements: molecular data nodes and interaction elements. Data nodes representing genes, RNA molecules, and proteins were retained as candidate graph nodes, whereas non-HGNC-resolvable elements—including microRNA-specific nodes—were excluded. Retained molecular nodes were initially extracted using their GPML pathway labels and available cross-reference identifiers. Before sample-level graph assembly, these nodes were canonicalized to HGNC gene symbols using available Ensembl, UniProt, or Entrez Gene cross-references queried through the `mygene`[97] Python package. Overall, this workflow ensured that each final graph node corresponded to a measured gene feature in the input transcriptomic expression matrix.

Next, edges were inferred from GPML interaction elements and added to the pathway graph. Each interaction was inspected for graphical points that referenced retained molecular nodes through GPML `GraphRef` attributes. Interactions between two retained molecular nodes were represented as undirected edges connecting the corresponding nodes. When an interaction referenced more than two retained molecular nodes, all pairwise edges among those nodes were added. The resulting HGNC-standardized pathway p was therefore represented as an undirected molecular interaction graph:

$$H_p = (V_p, E_p) \tag{1}$$

Here, V_p denotes the set of retained molecular nodes in pathway p, and E_p denotes the set of curated pathway interactions extracted from the GPML representation. Node labels in H_p were then canonicalized to HGNC gene symbols. Specifically, available Ensembl, UniProt, and Entrez Gene cross-references were queried using the `mygene` Python client for MyGene.info, and nodes that could not be mapped to valid HGNC symbols were removed. After canonicalization, pathway graphs were again filtered to retain only genes present in the processed mRNA expression matrix. Self-loops introduced by identifier collapsing were removed, and the resulting HGNC-standardized pathway graph was denoted as follows:

$$\tilde{H}_p = (\tilde{V}_p, \tilde{E}_p), \qquad \tilde{V}_p \subseteq G \tag{2}$$

where G denotes the set of measured genes retained in the processed transcriptomic expression matrix. Note that each processed pathway graph was cached by WPID and reused across samples, ensuring that pathway retrieval, GPML parsing, and HGNC canonicalization were performed once per pathway rather than repeatedly for every sample. After individual pathway graphs were constructed, all pathways selected for a given sample were merged into a single undirected molecular interaction graph for that patient. Overall, this procedure preserved WikiPathways-derived gene-gene connectivity while focusing the final graph on genes and pathways associated with the strongest patient-specific expression deviations.

4.5 Assembly of patient-specific pathway graphs

For each sample i, all processed pathway graphs corresponding to the selected pathway set P_i were merged into a single patient-specific graph. The resulting graph was denoted as follows:

$$\mathscr{G}_i = (V_i, E_i) = \bigcup_{p \in P_i} \tilde{H}_p \tag{3}$$

Here, $\mathscr{G}_i$ denotes the patient-specific graph for sample i, P_i denotes the set of curated biological pathways selected from WikiPathways for that sample, and $\tilde{H}_p$ denotes the HGNC-standardized[93] node and edge sets for pathway p. The node and edge sets of $\mathscr{G}_i$ were defined as the union of the node and edge sets from all selected pathway graphs:

$$V_i = \bigcup_{p \in P_i} \tilde{V}_p, \qquad E_i = \bigcup_{p \in P_i} \tilde{E}_p \tag{4}$$

Here, V_i and E_i denote the final node and edge sets of sample i, whereas $\tilde{V}_p$ and $\tilde{E}_p$ denote the HGNC-standardized[93] node and edge sets of individual pathway p. Genes appearing in multiple pathways were collapsed into a single node based on their HGNC symbol, while edges from all selected pathways were retained. This construction allowed shared genes to form bridges between pathway modules, producing an integrated sample-specific molecular interaction network. Self-loops were removed after merging, and each final graph was then used as the patient-level input for downstream graph neural network modeling. In this representation, nodes correspond to measured HGNC genes, edges correspond to curated WikiPathways interactions, and graph topology varies across patients according to the pathways selected from each patient's gene expression profile.

4.6 Graph-level dataset assembly

For each prediction task, patient-specific pathway graphs were paired with matched transcriptomic expression profiles and task labels before model training. Upon parsing the graph files, only samples with corresponding expression profile, clinical label, and patient-specific graph were retained. After sample alignment, task labels were integer-encoded and each patient graph was converted into a PyTorch Geometric `Data` object[98] for graph-level classification.

For a sample i, the graph object was represented as:

$$\mathscr{D}_i = (X_i, \mathbf{g}_i, \text{edge\_index}_i, y_i)\,. \tag{5}$$

Here, $X_i \in \mathbb{R}^{|V_i|\times 2}$ denotes the two-channel node feature matrix for graph $\mathcal{G}_i$, $\mathbf{g}_i$ denotes the integer-encoded gene identity for each node, edge_index$_i$ denotes the graph connectivity used for message passing, and y_i denotes the encoded sample-level class label. Graphs with fewer than two nodes or without valid edges were excluded from model training. Because the pathway graphs were treated as undirected for message passing, each edge was represented in both directions in the PyTorch Geometric edge index.

4.7 Node feature construction

Each graph node was assigned two features representing patient-specific expression and cohort-level co-expression. For the RARE-Seq and prostate cancer datasets, expression values were transformed using $\log(1+x)$ before feature construction, whereas the pan-cancer expression matrix was used on its provided scale.

For gene g in sample i, the patient-specific expression feature was calculated by standardizing expression values across genes within that sample:

$$z_{gi} = \frac{T(x_{gi}) - \mu_i}{\sigma_i + \varepsilon}, \tag{6}$$

where x_{gi} is the expression value of gene g, $T(\cdot)$ denotes the dataset-specific transformation, and μ_i and σ_i are the mean and standard deviation across genes in sample i.

The co-expression feature was derived from pairwise gene correlations across the cohort. Expression values were first standardized for each gene across samples:

$$\tilde{x}_{gi} = \frac{T(x_{gi}) - \mu_g}{\sigma_g + \varepsilon}, \tag{7}$$

where μ_g and σ_g are the mean and standard deviation of gene g across samples. Pairwise gene correlations were then calculated as:

$$r_{gh} = \frac{1}{N}\sum_{i=1}^{N} \tilde{x}_{gi}\tilde{x}_{hi}. \tag{8}$$

For each gene g, the co-expression feature was defined as the mean absolute correlation with its 50 most strongly correlated genes, excluding self-correlation:

$$c_g = \frac{1}{k}\sum_{h\in \mathrm{TopK}(g)} |r_{gh}|, \qquad k = 50. \tag{9}$$

The resulting node-feature vector was:

$$\mathbf{x}_{gi} = \begin{bmatrix} z_{gi} \\ c_g \end{bmatrix}. \tag{10}$$

Thus, each node contained a patient-specific expression feature and a cohort-level co-expression feature.

4.8 Gene identity encoding

Each gene was assigned an integer index and mapped to a trainable gene-identity embedding:

$$\mathbf{e}_g = \mathbf{E}[\mathrm{id}(g)], \tag{11}$$

where $\mathbf{E}$ is the embedding matrix and $\mathrm{id}(g)$ is the index assigned to gene g. This embedding preserved gene identity across patient graphs with different node sets.

The two node features were projected into a latent representation and combined with the gene-identity embedding:

$$\mathbf{h}_{gi}^{(0)} = f_{\mathrm{node}}\left(\mathrm{ReLU}(W_x\mathbf{x}_{gi} + \mathbf{b}_x) \,\|\, \mathbf{e}_g\right), \tag{12}$$

where $\|$ denotes concatenation. The expression projection and gene embedding each had 64 dimensions, and the resulting node representation had 128 dimensions.

4.9 Graph neural network architecture

GiG was implemented as a graph-level classifier with two message-passing layers. Four graph neural network backbones were evaluated: GCN, GraphSAGE, GIN and GATv2[28,99–101]. All backbones used the same node features, patient-specific graphs, pooling operation, and classification head.

The message-passing operation was expressed as:

$$\mathbf{H}_i^{(\ell+1)} = \mathrm{GNN}_\theta^{(\ell)}\left(\mathbf{H}_i^{(\ell)}, \mathrm{edge\_index}_i\right), \qquad \ell \in \{0,1\}, \tag{13}$$

where $\mathbf{H}_i^{(\ell)}$ contains the node representations for graph i at layer ℓ. Pathway interactions were represented as unweighted, undirected edges.

After message passing, node representations were aggregated by global mean pooling:

$$\mathbf{h}_{\mathscr{G}_i} = \frac{1}{|V_i|}\sum_{v\in V_i}\mathbf{h}_{vi}^{(2)}. \tag{14}$$

The resulting graph representation was passed through a classification head comprising a linear layer, ReLU activation, dropout, and a final output layer:

$$\hat{\mathbf{y}}_i = f_{\mathrm{head}}\left(\mathbf{h}_{\mathscr{G}_i}\right). \tag{15}$$

The output dimension was set to the number of classes in the corresponding prediction task.

4.10 Model training and class-weighted optimization

Class imbalance was addressed using class-weighted cross-entropy loss. For a task with C classes, the weight assigned to class c was calculated from the number of training samples n_c:

$$w_c = \frac{1/n_c}{\sum_{j=1}^{C} 1/n_j}C. \tag{16}$$

The training objective for a mini-batch $\mathscr{B}$ was:

$$\mathscr{L} = -\frac{1}{|\mathscr{B}|}\sum_{i\in\mathscr{B}} w_{y_i}\log\left(\frac{\exp(\hat{y}_{i,y_i})}{\sum_{c=1}^{C}\exp(\hat{y}_{i,c})}\right). \tag{17}$$

Models were optimized using AdamW[102] with a learning rate of 10^{-3}, weight decay of 10^{-4}, batch size of 32, and dropout probability of 0.2. Gradients were clipped to a maximum norm of 5.0. The checkpoint with the best validation performance was retained within each fold.

4.11 Cross-validation and performance evaluation

For the RARE-Seq and prostate cancer analyses, performance was assessed using stratified five-fold cross-validation. In each fold, one partition was reserved for testing. The remaining samples were divided into training and validation sets, with 25% assigned to validation. Model checkpoints were selected using validation performance and evaluated on the corresponding test fold. Results were summarized across the five test folds.

For the pan-cancer analysis, 10% of the samples were first reserved as a fixed stratified test set. The remaining samples were evaluated using stratified five-fold cross-validation. The model selected within each fold was applied to the fixed test set, and performance was summarized across the five trained models.

Performance was evaluated using accuracy, macro-F1, sensitivity, and specificity. Accuracy was defined as:

$$\mathrm{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\hat{y}_i = y_i\right), \tag{18}$$

where N is the number of evaluated samples and $\mathbb{I}$ is the indicator function. Macro-F1 was calculated as the unweighted mean of the class-specific F1 scores:

$$\mathrm{F1} = \frac{1}{C}\sum_{c=1}^{C}\frac{2\,\mathrm{Precision}_c\,\mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}. \tag{19}$$

Sensitivity and specificity were defined for each class as:

$$\mathrm{Sensitivity}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \qquad \mathrm{Specificity}_c = \frac{\mathrm{TN}_c}{\mathrm{TN}_c + \mathrm{FP}_c}. \tag{20}$$

For multiclass tasks, class-specific metrics were macro-averaged unless otherwise specified. Receiver operating characteristic curves, precision-recall curves, and confusion matrices were generated from the corresponding test predictions.

4.12 Random graph controls and graphlet analysis

To assess the contribution of pathway topology, each patient-specific graph was compared with Erdős–R'enyi and degree-preserving random controls. Erdős–R'enyi graphs were generated using the same nodes and number of edges as the corresponding pathway graph[89]. Degree-preserving controls were generated by rewiring edges while retaining the degree of each node[90]. The randomized graphs retained the original node features and clinical labels and were evaluated using the same model architecture and training procedure as the pathway-derived graphs. Local structural properties were compared using graphlet orbit counts calculated with ORCA[91].

5 Datasets

Unless otherwise specified, highly variable genes (HVGs) were selected based on variance across samples or cells within each dataset. All tabular inputs were standardized prior to modeling.

5.1 Circulating cell-free RNA profiling (RARE-Seq).

We evaluated GiG on the RARE-Seq circulating cell-free RNA (cfRNA) cohort, a multi-cancer liquid biopsy dataset designed to detect low-abundance tumor-derived transcripts in plasma[32]. This dataset contains 419 plasma cfRNA expression profiles from healthy individuals and patients with multiple cancer types, with available annotations for cancer status, tissue of origin, and clinical stage. Because cfRNA measurements contain tumor-derived signal diluted within a heterogeneous background of circulating RNA, this cohort provides a challenging setting for evaluating whether pathway-structured graph representations can recover weak but biologically meaningful disease signals. The input expression matrix was restricted to Ensembl gene identifiers. Ensembl version suffixes were removed, and identifiers were mapped to HGNC gene symbols using protein-coding annotations. Genes without valid HGNC symbols were excluded. To reduce the contribution of low-variance and noise-dominated features, the top 10,000 highly variable genes were selected based on variance across samples before patient-specific pathway graph construction. Three classification tasks were defined from the same processed expression and graph inputs. For binary cancer detection, samples were grouped as Healthy or Cancer, excluding categories that were not part of either class definition. For subtype classification, cancer samples were mapped to LUAD, LIHC, PAAD, and PRAD categories, with LUAD (TKI) samples grouped under LUAD. For stage classification, only cancer samples with available stage annotations were retained and classified as stage I, II, III, or IV. Patient-specific pathway graphs were then constructed from the most dysregulated genes in each sample and paired with the corresponding task labels for downstream graph-level classification.

5.2 Prostate cancer transcriptomic dataset.

We next evaluated GiG on a tissue-derived prostate cancer transcriptomic dataset containing 660 localized and metastatic prostate cancer samples[14]. In contrast to cfRNA profiling, tissue-based transcriptomic measurements capture more direct tumor-associated molecular programs and therefore represent a higher-signal setting for evaluating pathway-informed graph learning. This dataset was used to test whether patient-specific pathway graphs could distinguish clinically meaningful prostate cancer disease states while prioritizing genes involved in established prostate cancer signaling programs, including PI3K/AKT, MAPK, and Wnt-related pathways. The raw prostate cancer expression matrix was reformatted so that samples and genes were consistently aligned with the graph-construction pipeline. Gene symbols were already provided in HGNC-compatible format, so no additional gene identifier conversion was required. Genes were ranked by variance across samples, and the top 10,000 highly variable genes were retained before graph construction. Patient-specific pathway graphs were generated using the same WikiPathways-based pipeline applied to the other cohorts. The classification task was formulated as binary prediction of localized versus metastatic prostate cancer. Each sample was retained only when a matched expression profile, class label, and patient-specific pathway graph were available.

5.3 Pan-cancer transcriptomic dataset.

To evaluate GiG in a large multiclass setting, we used a pan-cancer transcriptomic dataset derived from The Cancer Genome Atlas (TCGA)[33,34]. This cohort spans diverse tumor types and tissue lineages from 8,315 patient samples, providing a stringent benchmark for testing whether pathway-informed graph representations can scale beyond binary classification to a high-dimensional multiclass prediction problem. Unlike the RARE-Seq and prostate cancer tasks, the pan-cancer task requires the model to distinguish many cancer types with partially overlapping transcriptional programs, making balanced classification and class-specific feature attribution particularly important. The pan-cancer input matrix was provided as a pre-aligned gene expression matrix and was used directly as the gene universe for graph construction. No additional HVG filtering was applied at this stage in order to preserve the aligned gene set across cancer types and avoid removing genes with potential lineage- or subtype-specific relevance. For each sample, genes with the strongest within-sample expression deviations were selected and mapped to WikiPathways to construct a patient-specific pathway graph. The task was formulated as 32-class cancer type classification. Each graph was paired with the corresponding cancer type label and used for graph-level model training and evaluation.

5.4 Urothelial carcinoma transcriptomic dataset (GSE32894).

To evaluate GiG across clinically distinct phenotypes within a single disease, we analyzed GSE32894[73], a urothelial carcinoma transcriptomic cohort comprising 308 fresh-frozen tumor resection specimens. Urothelial carcinoma is characterized by substantial molecular and clinical heterogeneity[74–76], thus this cohort provides an informative setting for testing whether patient-specific pathway graphs can recover clinically relevant disease phenotypes that reflect overlapping but distinct aspects of tumor biology. Raw microarray probe identifiers were first mapped to HGNC gene symbols using the corresponding Illumina platform annotation. Probes without valid gene mappings were excluded, and the most highly variable genes were retained before patient-specific pathway graph construction. Selected genes were processed through the same WikiPathways-based pipeline used for the primary cohorts, such that each tumor was represented by a graph containing its most transcriptionally dysregulated genes organized according to curated molecular interactions. Three complementary clinical classification tasks were constructed from the available annotations: tumor progression, tumor grade, and tumor stage. For each task, only samples with matched expression profiles, clinical annotations, and successfully constructed patient-specific graphs were retained for downstream classification.

5.5 Colon cancer transcriptomic dataset (GSE39582).

We evaluated GiG using GSE39582[103], a large colon cancer transcriptomic cohort originally developed to characterize molecular heterogeneity within colorectal malignancy. The dataset contains 585 expression profiles, including 566 primary colon cancers and 19 non-tumoral colorectal mucosa samples. The original study demonstrated that colon cancers comprise transcriptionally distinct molecular states with differences in genomic alterations, anatomical distribution, and clinical outcome[103], making this dataset well suited for evaluating whether pathway-structured representations can recover biologically meaningful phenotypes beyond conventional tumor classification. Affymetrix probe identifiers were mapped to HGNC gene symbols using the GPL570 platform annotation. Probes without valid HGNC mappings were removed, and redundant probes corresponding to the same gene were collapsed to generate a standardized gene-level expression matrix. The most highly variable genes were retained for pathway mapping and patient-specific graph construction, and samples lacking the clinical or molecular annotation required for a given task were excluded. Two classification tasks were evaluated: BRAF mutation status and tumor location. Patient-specific pathway graphs were paired with the corresponding task labels for graph-level model training and evaluation.

5.6 Gastric cancer transcriptomic dataset (GSE62254).

GSE62254[104] was used to evaluate GiG in gastric cancer, a clinically and molecularly heterogeneous malignancy for which biologically distinct subgroups have substantially different patterns of disease progression and prognosis. This dataset was generated by the Asian Cancer Research Group (ACRG) and contains expression profiles from 300 primary gastric tumors. The ACRG study identified molecularly distinct gastric cancer states associated with markedly different clinical behaviors. Processed Affymetrix expression values were mapped from probe identifiers to HGNC gene symbols using the GPL570 annotation. Probes without valid gene mappings were excluded, redundant probes were collapsed at the gene level, and the most highly variable genes were used to generate pathway-structured graphs for each sample. In this dataset, the classification task focused on the epithelial–mesenchymal phenotype (EMP) subgroup annotation, corresponding to the mesenchymal/EMT-associated disease state represented in the ACRG molecular classification. Evaluated samples all have matched expression profiles, EMP subgroup labels, and patient-specific graphs.

5.7 Tumor-educated platelet RNA-seq dataset (GSE68086).

To extend evaluation beyond tumor tissue and plasma cfRNA, we analyzed GSE68086[79], a blood-based tumor-educated platelet (TEP) RNA-sequencing dataset developed for pan-cancer liquid biopsy applications. This cohort contains 285 platelet

samples, including 228 TEP samples from patients with cancer and 57 samples from healthy individuals. The cancer cohort spans six malignancies - non-small cell lung cancer, colorectal cancer, pancreatic cancer, glioblastoma, breast cancer, and hepatobiliary carcinoma - providing substantial heterogeneity in tissue of origin. For GiG preprocessing, Ensembl identifiers were first harmonized with the HGNC gene vocabulary used for WikiPathways graph construction, excluding genes without valid mappings. Genes were then ranked according to variance across samples before constructing patient-specific pathway graphs. The downstream task was formulated as binary classification of cancer-associated platelets versus platelets from healthy individuals. This task provides an additional test of GiG's applicability to minimally invasive cancer detection and, together with the RARE-Seq analysis, examines whether pathway-informed graph learning can extract meaningful signals from different sources of circulating transcriptomic information.

References

1. Hanahan, D. Hallmarks of cancer: new dimensions. *Cancer discovery* **12**, 31–46 (2022).
2. Drier, Y., Sheffer, M. & Domany, E. Pathway-based personalized analysis of cancer. *Proc. Natl. Acad. Sci.* **110**, 6388–6393 (2013).
3. Steyaert, S. *et al.* Multimodal data fusion for cancer biomarker discovery with deep learning. *Nat. machine intelligence* **5**, 351–362 (2023).
4. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
5. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Sci. data* **10**, 67 (2023).
6. Muneer, A. *et al.* From classical machine learning to emerging foundation models: review on multimodal data integration for cancer research. *Artif. Intell. Rev.* **59**, 119 (2026).
7. Hao, M. *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat. methods* **21**, 1481–1491 (2024).
8. Smith, A. M. *et al.* Standard machine learning approaches outperform deep representation learning on phenotype prediction from transcriptomics data. *BMC bioinformatics* **21**, 119 (2020).
9. Milacic, M. *et al.* The reactome pathway knowledgebase 2024. *Nucleic acids research* **52**, D672–D678 (2024).
10. Szklarczyk, D. *et al.* The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research* **51**, D638–D646 (2023).
11. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic acids research* **51**, D587–D592 (2023).
12. Wang, T. *et al.* Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. communications* **12**, 3445 (2021).
13. Valous, N. A., Popp, F., Zörnig, I., Jäger, D. & Charoentong, P. Graph machine learning for integrated multi-omics analysis. *Br. journal cancer* **131**, 205–211 (2024).
14. Elmarakeby, H. A. *et al.* Biologically informed deep neural network for prostate cancer discovery. *Nature* **598**, 348–352 (2021).
15. Hartman, E. *et al.* Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis. *Nat. Commun.* **14**, 5359 (2023).
16. Ma, J. *et al.* Using deep learning to model the hierarchical structure and function of a cell. *Nat. methods* **15**, 290–298 (2018).
17. Lin, C.-H. & Lichtarge, O. Using interpretable deep learning to model cancer dependencies. *Bioinformatics* **37**, 2675–2681 (2021).
18. Jha, K., Saha, S. & Singh, H. Prediction of protein–protein interaction using graph neural networks. *Sci. Reports* **12**, 8360 (2022).
19. Gao, Z. *et al.* Hierarchical graph learning for protein–protein interaction. *Nat. Commun.* **14**, 1093 (2023).
20. Kesimoglu, Z. N. & Bozdag, S. Supreme: multiomics data integration using graph convolutional networks. *NAR Genomics Bioinforma.* **5**, lqad063 (2023).
21. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. methods* **11**, 333–337 (2014).

22. Pai, S. *et al.* netdx: interpretable patient classification using integrated patient similarity networks. *Mol. systems biology* **15**, MSB188497 (2019).

23. Huang, K. *et al.* A foundation model for clinician-centered drug repurposing. *Nat. Medicine* **30**, 3601–3613 (2024).

24. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).

25. Agrawal, A. *et al.* Wikipathways 2024: next generation pathway database. *Nucleic acids research* **52**, D679–D689 (2024).

26. Slenter, D. N. *et al.* Wikipathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic acids research* **46**, D661–D667 (2018).

27. Zhang, S., Tong, H., Xu, J. & Maciejewski, R. Graph convolutional networks: a comprehensive review. *Comput. Soc. Networks* **6**, 1–23 (2019).

28. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, vol. 30 (2017).

29. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).

30. Veličković, P. *et al.* Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

31. Liu, Z. & Zhou, J. Graph attention networks. In *Introduction to graph neural networks*, 39–41 (Springer, 2022).

32. Nesselbush, M. C. *et al.* An ultrasensitive method for detection of cell-free rna. *Nature* **641**, 759–768 (2025).

33. Weinstein, J. N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nat. genetics* **45**, 1113–1120 (2013).

34. Hoadley, K. A. *et al.* Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304.e6, DOI: 10.1016/j.cell.2018.03.022 (2018).

35. Bollobás, B. Random graphs. In *Modern graph theory*, 215–252 (Springer, 2011).

36. Campbell, C., Ruths, J., Ruths, D., Shea, K. & Albert, R. Topological constraints on network control profiles. *Sci. reports* **5**, 18693 (2015).

37. Tantardini, M., Ieva, F., Tajoli, L. & Piccardi, C. Comparing methods for comparing networks. *Sci. reports* **9**, 17557 (2019).

38. Sarajlić, A., Malod-Dognin, N., Yaveroğlu, Ö. N. & Pržulj, N. Graphlet-based characterization of directed networks. *Sci. reports* **6**, 35098 (2016).

39. Thombre, K. R., Gabhane, V., Gupta, K. R. & Umekar, M. J. Network pharmacology in the multi-omics era: uncovering novel therapeutic strategies. *Discov. Chem.* **3**, 111 (2026).

40. Diao, Y. *et al.* Heterogeneity-aware, multiscale annotation of shared and specific neurobiological signatures among major neurodevelopmental disorders. *Research* **9**, 1115 (2026).

41. Chongtham, A. *et al.* Neocortical tau propagation is a mediator of clinical heterogeneity in alzheimer's disease. *Mol. psychiatry* **30**, 4194–4213 (2025).

42. Kenigsbuch, M. *et al.* A shared disease-associated oligodendrocyte signature among multiple cns pathologies. *Nat. neuroscience* **25**, 876–886 (2022).

43. Ornago, A. M. *et al.* Shared and specific blood biomarkers for multimorbidity. *Nat. medicine* **32**, 736–745 (2026).

44. Conard, A. M., DenAdel, A. & Crawford, L. A spectrum of explainable and interpretable machine learning approaches for genomic studies. *Wiley Interdiscip. Rev. Comput. Stat.* **15**, e1617 (2023).

45. Vorperian, S. K., Dennis, L. M., Hupalowska, A., Rood, J. E. & Quake, S. R. Cell type inference in cell-free nucleic acid liquid biopsy. *Nat. Biotechnol.* 1–14 (2025).

46. Morlion, A. *et al.* Patient-specific alterations in blood plasma cfrna profiles enable accurate classification of cancer patients and controls. *Commun. Medicine* (2026).

47. Ma, L. *et al.* Liquid biopsy in cancer: current status, challenges and future prospects. *Signal transduction targeted therapy* **9**, 336 (2024).

48. Cabús, L., Lagarde, J., Curado, J., Lizano, E. & Pérez-Boza, J. Current challenges and best practices for cell-free long rna biomarker discovery. *Biomark. research* **10**, 62 (2022).

49. Ghosh, K. *et al.* The class imbalance problem in deep learning. *Mach. Learn.* **113**, 4845–4901 (2024).

50. Jackson, B. R. The dangers of false-positive and false-negative test results: false-positive results as a function of pretest probability. *Clin. laboratory medicine* **28**, 305–319 (2008).

51. Burt, T., Button, K., Thom, H., Noveck, R. & Munafò, M. R. The burden of the "false-negatives" in clinical development: Analyses of current and alternative scenarios and corrective measures. *Clin. translational science* **10**, 470–479 (2017).

52. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, vol. 70 of *Proceedings of Machine Learning Research*, 3319–3328 (PMLR, 2017).

53. Nawaz, F. Z. & Kipreos, E. T. Emerging roles for folate receptor folr1 in signaling and cancer. *Trends Endocrinol. & Metab.* **33**, 159–174 (2022).

54. Ju, Y., Fang, S., Liu, L., Ma, H. & Zheng, L. The function of the elf3 gene and its mechanism in cancers. *Life Sci.* **346**, 122637 (2024).

55. Saha, S. *et al.* Krt19 directly interacts with β-catenin/rac1 complex to regulate numb-dependent notch signaling pathway and breast cancer properties. *Oncogene* **36**, 332–349 (2017).

56. Bax, H. J. *et al.* Folate receptor alpha in ovarian cancer tissue and patient serum is associated with disease burden and treatment outcomes. *Br. journal cancer* **128**, 342–353 (2023).

57. Leung, F., Dimitromanolakis, A., Kobayashi, H., Diamandis, E. & Kulasingam, V. Folate-receptor 1 (folr1) protein is elevated in the serum of ovarian cancer patients. *Clin. biochemistry* **46**, 1462–1468 (2013).

58. Saloustros, E. *et al.* Cytokeratin-19 mrna-positive circulating tumor cells during follow-up of patients with operable breast cancer: prognostic relevance for late relapse. *Breast cancer research* **13**, R60 (2011).

59. Carpten, J. D. *et al.* A transforming mutation in the pleckstrin homology domain of akt1 in cancer. *Nature* **448**, 439–444 (2007).

60. Sinkala, M., Nkhoma, P., Mulder, N. & Martin, D. P. Integrated molecular characterisation of the mapk pathways in human cancers reveals pharmacologically vulnerable mutations and gene dependencies. *Commun. Biol.* **4**, 9 (2021).

61. Tanabe, S., Kawabata, T., Aoyagi, K., Yokozaki, H. & Sasaki, H. Gene expression and pathway analysis of ctnnb1 in cancer and stem cells. *World journal stem cells* **8**, 384 (2016).

62. Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215–1228 (2015).

63. Abida, W. *et al.* Genomic correlates of clinical outcome in advanced prostate cancer. *Proc. Natl. Acad. Sci.* **116**, 11428–11436 (2019).

64. Litwin, M. S. & Tan, H.-J. The diagnosis and treatment of prostate cancer: a review. *Jama* **317**, 2532–2542 (2017).

65. Shore, N. D. Current and future management of locally advanced and metastatic prostate cancer. *Rev. Urol.* **22**, 110 (2020).

66. Kim, E. S. *et al.* Potential utility of risk stratification for multicancer screening with liquid biopsy tests. *NPJ Precis. Oncol.* **7**, 39 (2023).

67. Passaro, A. *et al.* Cancer biomarkers: Emerging trends and clinical implications for personalized treatment. *Cell* **187**, 1617–1635 (2024).

68. Sicklick, J. K. *et al.* Molecular profiling of cancer patients enables personalized combination therapy: the i-predict study. *Nat. medicine* **25**, 744–750 (2019).

69. Colnot, S. *et al.* Liver-targeted disruption of apc in mice activates β-catenin signaling and leads to hepatocellular carcinomas. *Proc. Natl. Acad. Sci.* **101**, 17216–17221 (2004).

70. Lv, T. *et al.* Upr–cebpb–mif signaling links to macrophage polarization and an immunosuppressive tumor microenvironment in clear cell renal cell carcinoma. *Annals Surg. Oncol.* 1–17 (2026).

71. Zhang, Z. *et al.* Genome-wide crispr/cas9 screening for drug resistance in tumors. *Front. Pharmacol.* **14**, 1284610 (2023).

72. Moore, A. R. *et al.* Recurrent activating mutations of g-protein-coupled receptor cysltr2 in uveal melanoma. *Nat. genetics* **48**, 675–680 (2016).

73. Sjödahl, G. *et al.* A molecular taxonomy for urothelial carcinoma. *Clin. cancer research* **18**, 3377–3386 (2012).

74. Network, C. G. A. R. *et al.* Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315 (2014).

75. Chu, C. E. *et al.* Clinical outcomes, genomic heterogeneity, and therapeutic considerations across histologic subtypes of urothelial carcinoma. *Eur. urology* (2025).

76. Knowles, M. A. & Hurst, C. D. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. *Nat. reviews cancer* **15**, 25–41 (2015).

77. Yoshino, H. *et al.* Characterization of phgdh expression in bladder cancer: potential targeting therapy with gemcitabine/cisplatin and the contribution of promoter dna hypomethylation. *Mol. Oncol.* **14**, 2190–2202 (2020).

78. Zheng, H. *et al.* Tumor-derived exosomal bcyrn1 activates wnt5a/vegf-c/vegfr3 feedforward loop to drive lymphatic metastasis of bladder cancer. *Clin. translational medicine* **11**, e497 (2021).

79. Best, M. G. *et al.* Rna-seq of tumor-educated platelets enables blood-based pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer cell* **28**, 666–676 (2015).

80. Best, M. G., Wesseling, P. & Wurdinger, T. Tumor-educated platelets as a noninvasive biomarker source for cancer detection and progression monitoring. *Cancer research* **78**, 3407–3412 (2018).

81. Li, S. *et al.* The dynamic role of platelets in cancer progression and their therapeutic implications. *Nat. Rev. Cancer* **24**, 72–87 (2024).

82. Mócsai, A., Ruland, J. & Tybulewicz, V. L. The syk tyrosine kinase: a crucial player in diverse biological functions. *Nat. Rev. Immunol.* **10**, 387–402 (2010).

83. Suzuki-Inoue, K. Essential in vivo roles of the platelet activation receptor clec-2 in tumour metastasis, lymphangiogenesis and thrombus formation. *The J. Biochem.* **150**, 127–132 (2011).

84. Zheng, P. *et al.* Deciphering the molecular and clinical characteristics of trem2, hcst, and tyrobp in cancer immunity: A comprehensive pan-cancer study. *Heliyon* **10** (2024).

85. Karp, J. M. *et al.* Deconvolution of the tumor-educated platelet transcriptome reveals activated platelet and inflammatory cell transcript signatures. *JCI insight* **9**, e178719 (2024).

86. Montégut, L. *et al.* Acyl-coenzyme a binding protein (acbp)-a risk factor for cancer diagnosis and an inhibitor of immunosurveillance. *Mol. Cancer* **23**, 187 (2024).

87. Hacein-Bey-Abina, S. *et al.* Erythropoietin is a major regulator of thrombopoiesis in thrombopoietin-dependent and-independent contexts. *Exp. Hematol.* **88**, 15–27 (2020).

88. Siravegna, G., Marsoni, S., Siena, S. & Bardelli, A. Integrating liquid biopsies into the management of cancer. *Nat. reviews Clin. oncology* **14**, 531–548 (2017).

89. Erdős, P. & Rényi, A. On random graphs i. *Publ. Math.* **6**, 290–297 (1959).

90. Maslov, S. & Sneppen, K. Specificity and stability in topology of protein networks. *Science* **296**, 910–913 (2002).

91. Hočevar, T. & Demšar, J. A combinatorial approach to graphlet counting. *Bioinformatics* **30**, 559–565, DOI: 10.1093/bioinformatics/btt717 (2014).

92. Clough, E. *et al.* Ncbi geo: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic acids research* **52**, D138–D144 (2024).

93. Seal, R. L. *et al.* Genenames. org: the hgnc and pgnc resources in 2026. *Nucleic Acids Res.* **54**, D1098–D1107 (2026).

94. Hubbard, T. *et al.* The ensembl genome database project. *Nucleic acids research* **30**, 38–41 (2002).

95. Wu, C., Mark, A. & Su, A. I. Mygene. info: gene annotation query as a service. *bioRxiv* 009332 (2014).

96. Povey, S. *et al.* The hugo gene nomenclature committee (hgnc). *Hum. genetics* **109**, 678 (2001).

97. Wu, C., MacLeod, I. & Su, A. I. Biogps and mygene. info: organizing online, gene-centric information. *Nucleic acids research* **41**, D561–D565 (2013).

98. Fey, M. & Lenssen, J. E. Fast graph representation learning with pytorch geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds* (2019).

99. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations* (2017).

100. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations* (2019).

101. Brody, S., Alon, U. & Yahav, E. How attentive are graph attention networks? In *International Conference on Learning Representations* (2022).

102. Loshchilov, I. & Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations* (2019).

103. Marisa, L. *et al.* Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS medicine* **10**, e1001453 (2013).

104. Cristescu, R. *et al.* Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. *Nat. medicine* **21**, 449–456 (2015).