TRAPS: Treatment-Assignment Prediction via Pathway-informed Stratification

Abstract

Cancer treatment planning requires decisions across multiple clinical dimensions at once. Clinicians must determine whether a patient should receive targeted molecular therapy, radiation therapy, and whether they are likely to survive beyond six months. Existing pathway-informed deep learning models have been developed and tested in isolation, so it is unclear how they compare on a common footing. We present a harmonized benchmark for pathway-guided prediction of treatment assignment and short-term survival, evaluating three biologically informed architectures, BINN, GraphPath, and PATH, across five cancer cohorts drawn from The Cancer Genome Atlas, representing 2,622 patients encoded using Reactome pathway activity scores. We emphasise that the labels reflect the treatment a patient was recorded as receiving in TCGA, i.e. treatment exposure, rather than a measured response to therapy. Each model is trained jointly on all three clinical outcomes over a shared Reactome pathway representation, with each architecture retaining its original reference training protocol; to our knowledge this is the first study to treat pathway-structured deep learning as a combined treatment-assignment and short-term survival prediction problem. Our results show that, once every architecture is evaluated on identical stratified folds with matched uncertainty quantification (five repeated splits and paired-bootstrap tests), most inter-model differences fall within their 95% confidence intervals: the architecture rankings reported when these models are studied in isolation are largely not statistically resolved. The one consistent and significant effect is that the sparse-hierarchy model BINN is the strongest survival (OS) predictor, significantly outperforming both graph models on breast-cancer OS (ΔAUROC up to 0.14, $p \leq 0.01$) and leading on lung and prostate OS. For treatment-assignment prediction, targeted molecular therapy is best discriminated in the prostate cohort (AUROC $\approx$ 0.80 for all three models), yet no architecture significantly outperforms the others on any TMT cohort; radiation-therapy assignment is weakly predicted throughout, consistent with the hypothesis that its drivers are largely clinical rather than transcriptomic. Under a fair unified protocol, then, the choice among pathway-informed architectures matters far less than prior isolated results suggest, with short-term survival prediction the clearest point of differentiation.

CCS Concepts

• **Computing methodologies** → **Machine learning**; *Supervised learning by classification*; • **Applied computing** → **Computational biology**.

Keywords

biologically informed neural networks, graph attention, graph transformer, Reactome, ssGSEA, breast cancer, treatment-assignment prediction, multi-label classification, systems immunology, interpretable deep learning

1 Introduction

Cancer treatment planning is a difficult clinical problem because one patient may need several different kinds of decisions at the same time. A clinician may need to decide whether a patient should receive a targeted molecular therapy, whether radiation therapy should be used, and whether the patient is likely to survive beyond a short-term clinical window. These three questions are related, but they are not the same question. Targeted molecular therapy aims to block specific molecular drivers of cancer growth [29]. Radiation therapy uses ionising radiation to damage tumour cells [2]. Short-term survival prediction is often used to support risk stratification and care planning in oncology [31, 37]. A useful model should therefore learn these outcomes as separate but connected prediction tasks, instead of forcing them into one single label.

Moreover, the clinical reality is that these decisions are often made sequentially or concurrently, with information from one assessment influencing another, which further motivates a multi-task learning formulation that can capture shared and task-specific molecular signatures across outcomes.

Gene expression data can help with this problem because it captures which genes are active in each tumour sample. However, raw gene expression has thousands of genes, and it can be hard to understand what a model learns from such a large input. A more interpretable approach is to summarize gene expression into biological pathways. Gene set enrichment analysis was introduced to connect gene expression patterns with known biological processes [39], and single-sample GSEA makes it possible to estimate pathway activity for each individual sample [1]. In this paper, we use Reactome, a curated pathway knowledgebase, to convert each patient into a vector of pathway activity scores [11]. This representation is easier to connect back to biology because each feature corresponds to a known pathway rather than an isolated gene.

We formulate our problem as a multi-label prediction task. Given one patient's Reactome pathway activity profile, the model predicts three binary outcomes: targeted molecular therapy (TMT), radiation therapy (RT), and overall survival of at least 180 days (OS $\geq$ 180 d). This setup is important because a patient can belong to any combination of these outcomes. For example, one patient may receive targeted therapy but not radiation, while another may receive radiation and also survive beyond 180 days. By using three output heads, the model can learn which pathway patterns are most useful for each clinical question.

An important corollary of this formulation is that the three heads share a common biological substrate: genes that modulate targeted-drug sensitivity, radioresistance, and short-term survival are all expressed through the same transcriptome and aggregated into the same Reactome pathway activity vector. Joint training therefore permits gradient signals from a better-powered head to regularise the shared representation, which is particularly valuable in cohorts where one label is sparse.

Recent work has shown that biological knowledge can be built directly into deep learning models [47, 49, 55]. P-NET used pathway

structure to create a neural network for prostate cancer discovery [8]. BINN extended this idea by using biologically informed network connections to support pathway-level interpretation [13, 22]. GraphPath models pathways as nodes in a graph and uses graph attention to learn relationships between pathways [6, 16, 20, 21, 27, 41, 48]. PATH uses a graph transformer [7, 24, 50] to represent pathway interactions for cancer prognosis [14]. These models are promising, but they were developed and evaluated under different settings. Their datasets, labels, input features, and evaluation tasks are not the same, so it is difficult to know whether one architecture is generally better or whether each one works best for different clinical outcomes.

In this work, we compare these pathway-informed models under one common pipeline. We use gene expression and clinical data from The Cancer Genome Atlas (TCGA), accessed through the UCSC Xena platform [12, 44]. We study five solid-tumour cohorts: breast, lung, prostate, head and neck, and thyroid cancer. For every cohort, we use the same preprocessing strategy: gene expression is converted into Reactome pathway activity scores, clinical records are converted into the three binary labels, and the same multi-label learning objective is used for all models. This makes the comparison more direct because the input representation and the learning objective are shared across models. We stress, however, that each architecture retains its original reference training protocol, including its optimiser and hyperparameter settings, and for PATH, its own gene-membership pathway filter (Section 4); the benchmark is therefore harmonized at the level of data, representation, and evaluation splits rather than being a fully controlled head-to-head comparison, and cross-model rankings should be read with this caveat in mind.

Our goal is not only to report which model has the highest score, but also to understand when each kind of biological structure helps. BINN represents Reactome as a sparse hierarchy. GraphPath represents pathways as an attention-based graph. PATH uses a transformer-style graph model with pathway interactions. By testing these approaches on the same task, we ask a simple question: do different pathway-based architectures help with different therapy and survival outcomes? This question matters because label distributions are different across cancer types. A cohort with many patients receiving targeted therapy may not behave like a cohort where targeted therapy is rare. Therefore, the best model may depend on both the phenotype being predicted and the cancer cohort being studied.

The main contributions of this paper are as follows:

- We build a harmonized TCGA benchmark for pathway-based prediction of TMT, RT, and OS $\geq$ 6 m across five solid-tumour cohorts, sharing a common input representation and learning objective across architectures.
- We use the same Reactome and ssGSEA-based pathway representation for every patient, making the input space consistent across cohorts and models.
- We adapt and compare three pathway-informed deep learning architectures: BINN, GraphPath, and PATH.
- We show that model performance is phenotype-specific, meaning that the strongest architecture can change depending on whether the target is therapy assignment or short-term survival.

2 Related Work

Pathway-informed architectures, evaluated in isolation: The three models we benchmark were each introduced and validated on their own data, task, and protocol, which is precisely what makes a head-to-head comparison difficult. P-NET pioneered hard-wiring the Reactome hierarchy into a sparse network and was evaluated on a *single* task - metastatic vs. primary classification in prostate cancer—using its own train/test split and reporting AUROC/AUPRC on that one cohort [8]. BINN generalised this sparse-hierarchy idea for proteomic biomarker discovery and was assessed mainly through interpretability and stability of its layer-wise attributions rather than through cross-cohort predictive benchmarking [13, 22]. GraphPath reframed pathways as nodes in a graph-attention network and was reported on cancer-specific prognosis tasks with attention-derived pathway importance, again on curated cohorts chosen by its authors [6, 27, 48]. PATH introduced a graph-transformer with Jaccard-weighted edges and positional encodings and was evaluated on prognosis endpoints under its own gene-membership pathway filter [14, 24, 50]. Related pathway- and graph-based survival models (e.g., Cox-pathway and GNN-survival families) follow the same pattern of single-model, single-protocol reporting [23, 28, 35, 45, 54].

The benchmarking gap: Across these studies the datasets, label definitions, input features, splits, and evaluation metrics differ, so the published numbers are not commensurable: a higher AUROC reported for one architecture may reflect an easier cohort, a different endpoint, or a more favourable split rather than a better inductive bias. To our knowledge no prior work places BINN-, GraphPath-, and PATH-style architectures on a *single* pipeline—identical cohorts, an identical Reactome/ssGSEA input representation, identical stratified folds, and matched uncertainty quantification—so that differences can be attributed to architecture rather than to protocol. General surveys of biologically informed and graph neural networks call for exactly this kind of controlled comparison but do not themselves supply one [47, 49, 55]. This paper fills that gap: we hold data, representation, and evaluation splits fixed, add five-repeat confidence intervals and paired-bootstrap significance tests (Section 4), and report which single-model rankings survive a fair, shared-protocol re-evaluation.

3 Data

3.1 Data Curation

We curated a diverse gene expression profiles and clinical phenotype data from the UCSC Xena browser [12], an open-access platform providing harmonised large-scale cancer genomic datasets. Drawing from this resource, we selected five solid-tumour cohorts from The Cancer Genome Atlas [40, 44]: breast, lung, prostate, head and neck, and thyroid cancer. Together, these cohorts encompass a broad spectrum of biologically and clinically distinct tumour types, providing a representative foundation for assessing model generalisability across diverse cancer contexts.

3.2 Task Formation

Building on these cohorts, we defined three binary prediction tasks per patient to support clinically grounded evaluation. We stress at the outset that the therapy labels are derived from the treatment

a patient is *recorded as having received* in the TCGA clinical files, i.e. they capture treatment exposure or assignment, not a measured response of the tumour to that treatment. The tasks below should therefore be read as predicting which treatments were administered and whether the patient survived a short-term window, given the tumour's pathway activity profile.

Targeted Molecular Therapy (TMT): Classifying whether a patient was recorded as receiving compounds that selectively inhibit molecular drivers of tumour growth (e.g., classifying whether a breast cancer patient received a HER2-directed targeted compound) [29]. The label is set to 1 if the patient's drug_name or pharmaceutical-therapy records contain at least one agent mapped to a targeted-therapy class, and 0 otherwise.

Radiation Therapy (RT): Identifying whether a patient was recorded as undergoing ionising radiation treatment designed to induce lethal DNA damage in tumour cells while sparing adjacent healthy tissue (e.g., identifying whether a head and neck cancer patient received radiotherapy as part of their primary treatment course) [2]. The label is set to 1 if a radiation-therapy record is present for the patient and 0 otherwise.

Overall Survival ≥ 6 Months (OS ≥ 6 m): Determining whether a patient survived beyond the clinically established six-month (180-day) threshold, adopted as a surrogate of short-term prognosis (e.g., determining whether a lung cancer patient remained alive at least 180 days following diagnosis) [31, 37]. We formulate overall survival as a binary task (≥ 6 vs < 6 months) for short-term prognostic stratification in a multi-task setting. The label is computed from the TCGA `OS`/`OS.time` fields: a patient with a death event (`OS`= 1) before 180 days is labelled 0, and a patient whose last observed follow-up (`OS.time`) is ≥ 180 days is labelled 1. Because we treat the 180-day mark as a fixed classification threshold rather than a time-to-event outcome, patients who are censored (alive at last contact) before 180 days cannot be assigned an unambiguous label and are excluded from the OS head. This design does not model censoring or full time-to-event dynamics, so OS results reflect short-term risk classification rather than a full survival analysis; we return to this limitation in the Discussion.

This multi-task formulation acknowledges that a patient's treatment assignment and short-term prognosis are intimately connected through the underlying tumour biology, yet they require distinct predictive signals that may not be equally captured by any single architectural prior.

3.3 Preprocessing

We aggregated gene expression data into biological pathways using the Reactome database [11], retaining 1,706 curated pathways. Each pathway contains between 10 and 1,000 genes per sample. To quantify the biological activity within each retained pathway, scores were subsequently computed via single-sample gene set enrichment analysis (ssGSEA) [1, 39] and normalised to $[0, 1]$. In parallel, we systematically encoded clinical outcomes as binary labels corresponding to the three prediction tasks defined above: TMT, RT, and OS ≥ 6 m. Finally, pathway scores and outcome labels were merged using The Cancer Genome Atlas patient identifiers, retaining only complete samples for downstream analysis.

Table 1: Per-Cohort Sample Counts and Positive Prevalence of Each Tumor Phenotype Task Across All Five Cohorts

Cancer Type	*n*	TMT *n* (%)	RT *n* (%)	OS ≥ 6 m *n* (%)
Breast	618	571 (92%)	345 (56%)	462 (75%)
Prostate	496	54 (11%)	61 (12%)	480 (97%)
Head & Neck	429	150 (35%)	274 (64%)	395 (92%)
Thyroid	109	7 (6%)	61 (56%)	108 (99%)
Lung	970	311 (32%)	124 (13%)	863 (89%)
Total	**2,622**	**1,093 (42%)**	**865 (33%)**	**2,308 (88%)**

TMT = targeted molecular therapy; RT = radiation therapy; OS = overall survival.

4 Models

4.1 Model Selection

We evaluate three biologically informed deep learning architectures: **BINN**, **GraphPath**, and **PATH**. Each encodes a progressively richer notion of pathway structure, allowing us to ask how much the structural prior matters for treatment-assignment prediction. All other experimental conditions are held fixed across models, so any performance differences can be attributed to architectural choice alone.

The progression from BINN's strict hierarchical wiring through GraphPath's lateral connectivity to PATH's weighted graph transformer therefore permits a systematic investigation of how increasingly flexible structural priors affect predictive performance across diverse clinical endpoints.

All three models receive one standardised ssGSEA score per Reactome-matched pathway (1,706 pathways in total) and are trained end-to-end with a multi-task class-weighted binary cross-entropy objective over three prediction heads corresponding to TMT, RT, and OS ≥ 6 m. The architectures differ only in how biological knowledge is encoded. BINN wires its layers to follow Reactome parent-child relationships exclusively. GraphPath extends this by connecting pathways laterally as well as hierarchically within a graph attention network. PATH goes further by weighting edges with continuous Jaccard similarity and incorporating Laplacian positional encodings and edge-conditioned attention into a Graph Transformer. These three architectures represent a natural spectrum of structural commitment: BINN imposes hard biological constraints through sparse masked weights with no lateral pathway connections, GraphPath relaxes this to a soft lateral message-passing scheme via binary graph adjacency, and PATH goes furthest by learning continuous edge weights from gene-set overlap while incorporating global positional structure through Laplacian encodings. Holding all other variables constant allows any residual performance differences to be attributed to this structural axis rather than to preprocessing choices.

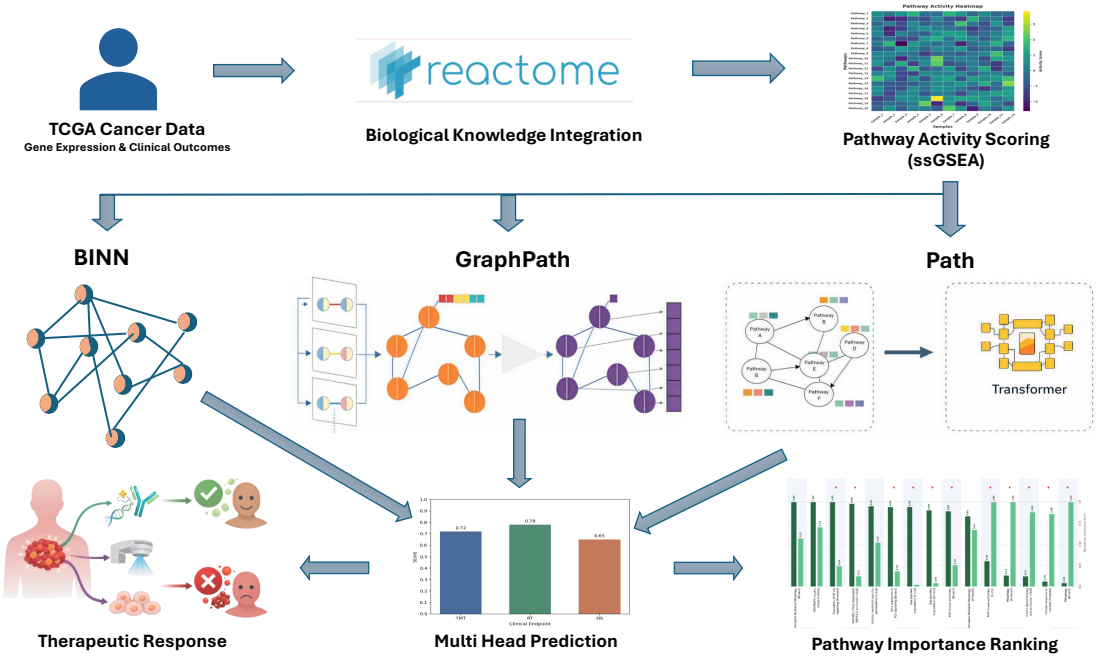

Figure 1: Workflow of the proposed method

4.2 Evaluation Metrics and Protocol

We report five complementary test-set metrics per clinical head: AUROC and AUPRC (both threshold-free), the positive-class F_1 and accuracy at the validation-tuned threshold, and the confusion-matrix counts; their formal definitions and the full per-model architecture equations are given in Appendix A. All three models minimise the same per-head class-weighted binary cross-entropy objective, and each keeps its reference optimiser and schedule (Appendix C).

4.2.1 Unified splits and evaluation protocol: To enable a controlled comparison, all three models share a single stratified 80/10/10 train/validation/test split per cohort, generated by a common splitter seeded identically across architectures. Consequently, at a fixed seed every model is trained and evaluated on *byte-identical* folds for a given cohort. To quantify uncertainty, we repeat the entire sweep over five seeds ($s \in \{0, 1, 2, 3, 4\}$), which yields five independent stratified splits per cohort, and report the mean and 95% confidence interval of each test metric across these repeats. In addition, because the models are aligned sample-for-sample on identical test folds, we compare architectures with a *paired bootstrap* over held-out samples (10^4 resamples): for each pair of models we recompute the AUROC difference on each resampled test set and report its 95% CI and a two-sided bootstrap p-value (Table 4). This directly addresses the need for repeated evaluation, confidence intervals, and statistical comparison under a shared split.

This evaluation protocol is deliberately conservative: by requiring models to be compared on identical folds with uncertainty estimates, we guard against the common pitfall of reporting performance on arbitrarily favourable splits. The paired bootstrap, in particular, accounts for the correlation between predictions made on the same test samples—a feature that standard independent t-tests would ignore—and therefore provides a more honest assessment of whether observed differences reflect genuine architectural advantages rather than sampling variability.

5 Results

Performance was evaluated across three clinical prediction heads TMT, RT, and OS using F1, AUPRC, and AUROC metrics on five TCGA solid-tumour cohorts. Unless otherwise noted, all reported numbers are the mean over five repeated stratified splits (seeds 0–4) evaluated on the unified 80/10/10 protocol, so that at each seed the three architectures see identical held-out folds. Table 2 reports the per-cohort, per-head test AUROC as mean [95% CI] across these repeats, and Table 4 reports paired-bootstrap significance tests for every pairwise model contrast. The per-head grouped-bar benchmark presented in Fig. 2 (three heads × five cohorts × three models) makes one thing immediately clear: no single architecture dominates uniformly. Direct comparison with the originally reported results of BINN, GraphPath and PATH is not appropriate because the models were evaluated under a different prediction task, feature representation, cohort composition and endpoint definition. Two patterns stand out once uncertainty is accounted for: (i) many head-to-head gaps have 95% CIs that include zero, so the ranking on those cells is not statistically resolved, and (ii) the differences that survive the bootstrap are concentrated in the TMT head, consistent with therapy-assignment outcomes being more architecture-sensitive than survival.

Table 2: Held-out test AUROC (mean [95% CI] over 5 identical-split repeats) for the three architectures on every cohort and clinical head. At a fixed repeat all three models are evaluated on byte-identical 80/10/10 stratified folds.

Cohort	Head	BINN	GraphPath	PATH
Breast	TMT	0.52 [0.40,0.65]	0.57 [0.52,0.63]	0.50 [0.43,0.57]
	RT	0.53 [0.47,0.59]	0.58 [0.47,0.69]	0.51 [0.41,0.62]
	OS	0.66 [0.60,0.72]	0.56 [0.50,0.61]	0.53 [0.49,0.58]
Lung	TMT	0.54 [0.53,0.56]	0.49 [0.47,0.51]	0.48 [0.43,0.54]
	RT	0.55 [0.47,0.63]	0.50 [0.49,0.52]	0.54 [0.47,0.61]
	OS	0.57 [0.51,0.62]	0.49 [0.44,0.55]	0.52 [0.47,0.58]
Prostate	TMT	0.80 [0.70,0.90]	0.79 [0.70,0.88]	0.82 [0.72,0.92]
	RT	0.69 [0.50,0.88]	0.62 [0.42,0.82]	0.66 [0.53,0.79]
	OS	0.73 [0.63,0.82]	0.66 [0.53,0.78]	0.45 [0.27,0.63]
Head & Neck	TMT	0.62 [0.55,0.70]	0.58 [0.53,0.63]	0.57 [0.49,0.66]
	RT	0.63 [0.58,0.69]	0.56 [0.49,0.63]	0.53 [0.46,0.60]
	OS	0.70 [0.61,0.78]	0.67 [0.49,0.86]	0.70 [0.50,0.91]
Thyroid	TMT	0.64 [0.24,1.04]	0.71 [0.50,0.93]	0.76 [0.48,1.04]
	RT	0.31 [0.12,0.51]	0.37 [0.13,0.60]	0.34 [0.23,0.45]
	OS	0.20	0.10	0.30

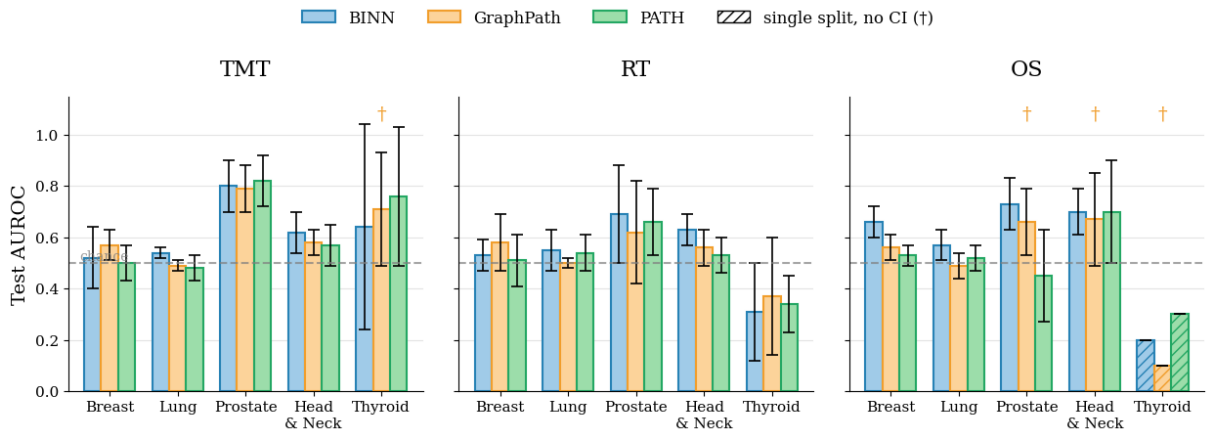

Figure 2: Test AUROC across cohorts and clinical heads, averaged over five identical-split repeats, with 95% CIs from Table 2. Colours follow the colourblind-safe Okabe–Ito palette: BINN blue, GraphPath orange, and PATH green [30, 46]. The dashed line shows chance performance (0.50). Hatched bars (†) mark cells that are single-split or underpowered and therefore lack a reliable CI, as described in Table 4.

5.1 Comparative Performance Across Clinical Endpoints

5.1.1 Evaluation of BINN:. BINN is the strongest survival predictor and the only architecture whose advantage is statistically resolved under the unified protocol. On the OS head it attains the highest mean test AUROC on the breast (0.66), lung (0.57), and prostate (0.73) cohorts (Table 2), and the paired-bootstrap tests confirm that its breast-OS advantage over both GraphPath (ΔAUROC 0.10, p=0.010) and PATH (ΔAUROC 0.14, p=0.001) is significant (Table 4). We attribute this to multi-depth auxiliary supervision, which propagates loss signals through every layer of the Reactome hierarchy and reinforces broad transcriptional survival signals distributed across pathway levels. The perfect thyroid OS-F1 seen in single-split runs does *not* survive repeated evaluation: with only a single OS-negative patient in the cohort, thyroid OS is degenerate and its estimates are flagged as underpowered (Table 4).

On the TMT head BINN is competitive but not dominant, reaching AUROC 0.62 on head & neck and 0.80 on prostate, statistically indistinguishable from the other two models on every cohort. On RT, BINN records the best head & neck AUROC (0.63), a significant margin over GraphPath (ΔAUROC 0.06, p=0.013); elsewhere RT differences are small and within confidence intervals.

5.1.2 Evaluation of PATH:. Under the unified protocol PATH does not establish a statistically significant advantage on any head. Its TMT AUROC is comparable to the other models on prostate (0.82 vs. 0.80/0.79; differences not significant) and lower on breast (0.50) and lung (0.48). The apparent head & neck TMT edge reported in isolated evaluation does not reproduce here (AUROC 0.57, statistically tied with BINN and GraphPath). On OS, PATH is the weakest model on breast (0.53) and prostate (0.45) and is significantly outperformed by BINN on breast OS (Table 4); it is only level with the others on head & neck OS.

On the RT head all three models converge, with no significant differences on breast, lung, or prostate. The dense Jaccard-weighted graph, though biologically motivated, does not translate into better discrimination on these cohorts, and its weaker pathway-level interpretability is not offset by a predictive gain.

5.1.3 Evaluation of GraphPath: GraphPath is broadly competitive with the other two architectures but is rarely separable from them: it is not the significant winner on any cohort/head. On TMT it is marginally strongest on breast (0.57) and thyroid (0.71), yet none of these gaps are statistically significant, and on RT it is significantly *beaten* by BINN on head & neck (Table 4). On OS it tracks the other models closely, trailing BINN significantly on breast.

The single-split "0.92 prostate TMT" result that stood out when GraphPath was evaluated in isolation does *not* hold under the unified protocol: across five identical-fold repeats its prostate TMT AUROC is 0.79 [0.70, 0.88], statistically indistinguishable from BINN (0.80) and PATH (0.82). This reversal is the kind of split-dependent artefact that a repeated-split evaluation exposes, and it is why cross-model claims from separately tuned splits should be treated with caution.

5.2 Cross-Cancer Subtype Performance

As shown in Fig. 3 and Table 2, performance variability across cohorts is more head- and cohort-dependent than model-dependent. On the TMT head, discrimination is driven overwhelmingly by the cohort rather than the architecture: all three models reach AUROC ≈ 0.80 on prostate—the most targetable-driver-defined cohort—while collapsing toward chance on breast and lung, and none of the inter-model TMT contrasts is statistically significant in any cohort (Table 4). The thyroid TMT cell, though numerically high for PATH, rests on only seven positive cases and carries very wide intervals.

On the RT head, the models perform similarly across cohorts and RT is weakly predicted overall; the only significant separation is at head & neck, where BINN exceeds GraphPath. On the OS head, BINN is consistently at or above the other models, significantly so on breast, while prostate and thyroid OS are too imbalanced to support reliable ranking. Across all three heads, the gaps that survive repeated evaluation and paired bootstrapping are concentrated in OS (favouring BINN) rather than in the treatment-assignment heads, indicating that - contrary to what isolated single-split results suggested - survival prognosis, not therapy assignment, is where architectural choice most clearly matters.This cohort-driven variability is itself informative: prostate TMT performance is uniformly high across models, reflecting known androgen-driven signatures, while the unstable thyroid TMT estimates highlight the value of repeated evaluation. The OS results show a consistent BINN advantage that survives bootstrapping, suggesting that short-term survival captures broad hierarchical signals BINN is designed to exploit. Therapy-assignment tasks, by contrast, show statistical parity across architectures, implying they depend on more focal signals or clinical factors beyond transcriptomics.

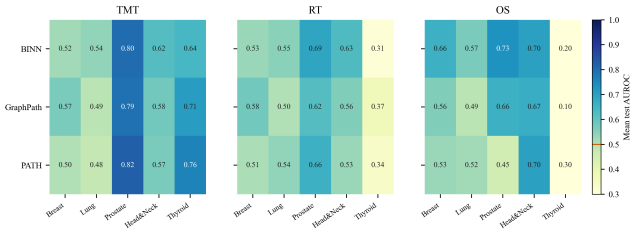

Figure 3: Mean test AUROC per (model, cohort) cell for each clinical head, using a perceptually uniform colourblind-safe sequential map; darker cells indicate higher AUROC and the colourbar marks chance (0.50). Every cell is annotated with its value so colour is never the sole channel. Values match Table 2.

5.3 Prediction Reliability Assessment

All three models are evaluated on the same unified 80/10/10 stratified split per cohort, so the confusion-matrix counts below are computed on identical held-out folds; the counts shown in Fig. 4 correspond to a single representative seed, while the mean±CI metrics of Table 2 and the paired-bootstrap tests of Table 4 summarise reliability across all five repeats. On the TMT head, all three models recover essentially every TMT-positive case (FN = 0 for all three; TP = 50 of 50), so recall is saturated and low FN counts confirm that pathway-level features encoding targetable molecular drivers are captured [29]. The difficulty lies entirely in the small TMT-negative subgroup: with only 12 negatives in the fold, BINN and GraphPath assign every patient to the positive class (TN = 0, FP = 12) while PATH recovers a single true negative (TN = 1). The elevated FP counts therefore reflect the dominant-positive prevalence (81% positive in this fold, 92% cohort-wide) rather than a systematic prediction failure.

The RT head has the most balanced label distribution (35 positives, 27 negatives) and is the most demanding setting for threshold-based prediction. Here the models diverge in how they treat negatives: GraphPath collapses to predicting RT for every patient (TN = 0, FP = 27), whereas BINN and PATH recover a comparable share of true negatives (TN = 6 and 7) at the cost of a few false negatives (FN = 5 and 7). This is consistent with RT being only weakly separable from pathway activity (Table 2): no model attains both high specificity and high recall on this head.

On the OS head (47 positives, 15 negatives), BINN and GraphPath achieve perfect recall by predicting survival for every patient (FN = 0, TN = 0), while PATH trades a small amount of recall (FN = 5) for two true negatives. False-negative counts are lowest on the OS and TMT heads, consistent with survival- and targeted-therapy-associated pathway signatures being the most consistently learned signal; the near-universal OS prevalence, however, means the confusion matrix alone cannot separate the models. It is the repeated-split AUROC (Table 2) and the paired-bootstrap tests (Table 4) that are decisive here, and both identify BINN as the strongest OS predictor [15, 19, 23, 28, 35, 42, 43, 45, 51, 54].

These patterns highlight the practical value of the unified evaluation protocol. Without repeated splits and ranking-based metrics, confusion matrices alone can easily mislead, potentially suggesting spurious differences or inflating confidence in threshold-dependent performance. The RT head, with its balanced distribution, offers the most rigorous test of discriminative capacity. Yet even in this favourable setting, no model manages to recover true negatives without sacrificing recall, confirming that radiation assignment is only weakly reflected in pathway activity profiles. For the TMT and OS heads, the story is different: saturating recall is largely uninformative and reflects label prevalence rather than genuine predictive power. Together, these observations underscore a broader principle: reliable model assessment in clinically imbalanced data demands careful consideration of metric choice, class prevalence, and statistical uncertainty—not a single confusion matrix. The full prediction breakdown is summarised in Fig. 4.

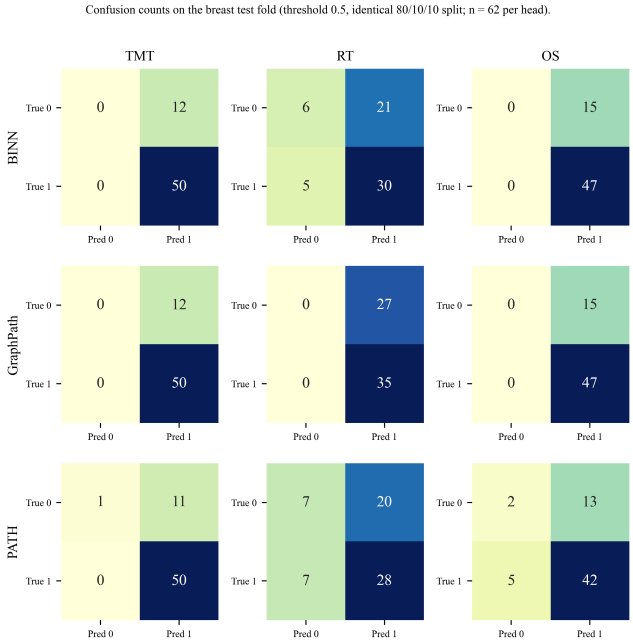

Figure 4: Confusion-matrix counts on the breast test fold at the 0.5 threshold, for a single representative seed. All three models are evaluated on the *same* 80/10/10 stratified split, so every panel sums to the same per-head test-set size ($n = 62$: TMT 50/12, RT 35/27, OS 47/15 positive/negative). Each row is a model, each column a clinical head; cells use the same colourblind-safe sequential map as Fig. 3 and are annotated with counts.

5.4 Pathway Importance Analysis and Biological Interpretation

Pathway importance scores from BINN (gradient × input attribution) and GraphPath (attention-derived importance) were compared across the TMT, RT, and OS heads to assess whether the two architectures converged on similar biological mechanisms. PATH was excluded because it does not generate pathway-level importance scores. Because the two architectures produce attributions on different and non-equivalent numerical scales, each model's scores were first min–max normalized to $[0, 1]$ within each head so that only the relative ranking of pathways is compared; the analysis is therefore qualitative and hypothesis-generating rather than a quantitative equivalence test. Pathways whose normalized scores differed by less than 0.50 were treated as convergent and prioritized for biological validation, whereas larger differences (>0.50; red arrows in Fig. 5) were flagged as potential architecture-specific effects. The 0.50 cut-off is a heuristic for surfacing candidates for manual review, not a calibrated significance threshold, and because it operates on min–max rescaled magnitudes rather than on the models' rankings of pathways, it should not be read as a measure of statistical agreement between the two attribution methods; the convergent pathways below should be read as literature-supported hypotheses rather than validated findings.

For the TMT head, the models emphasized different layers of the same oncogenic process, with BINN highlighting upstream transcriptional and inflammatory regulation and GraphPath focusing on downstream translational maintenance. Despite these differences, both converged on **ERK/MAPK Targets** [38] in Lung cancer, suggesting a shared biologically meaningful signal.

For the RT head, the strongest convergence was observed for **Receptor-Mediated Mitophagy** [53] [18] and **ERK/MAPK Targets** [3]. Both pathways have established links to radioresistance through mechanisms involving cellular stress adaptation, DNA repair, and survival signaling. Additional convergence on **VEGFR2-Mediated Vascular Permeability** [10] further supports the role of tumour microenvironment and vascular signaling in radiation response.

The OS head showed the highest agreement between the two models. Both architectures consistently ranked **ERK/MAPK Targets** [52] [9] among the most important pathways in both Lung and Head & Neck cohorts, making it the strongest convergent signal identified in this study. Its consistent appearance across multiple cohorts and clinical endpoints provides strong support for its biological relevance. In contrast, pathways related to apoptosis and metabolism showed lower agreement between the models, suggesting architecture-dependent interpretations that warrant further investigation. The convergence on ERK/MAPK signalling across multiple heads and cohorts is noteworthy not only for its biological plausibility but also because it emerges from two fundamentally different attribution mechanisms: gradient-based sensitivity in BINN and attention-derived importance in GraphPath. This cross-method agreement strengthens the case that the ERK/MAPK pathway is a genuinely informative signal for both therapy assignment and prognosis, rather than an artefact of a particular attribution technique. Conversely, the divergence observed for apoptosis and metabolism pathways highlights a limitation of the current approach. Without ground-truth biological validation, it remains unclear whether these discrepancies reflect genuine architectural differences in how each model represents pathway interactions, or merely noise introduced by the different attribution frameworks. Future work incorporating perturbational experiments or causal inference methods would help distinguish between these possibilities and refine the biological interpretability of pathway-informed models.

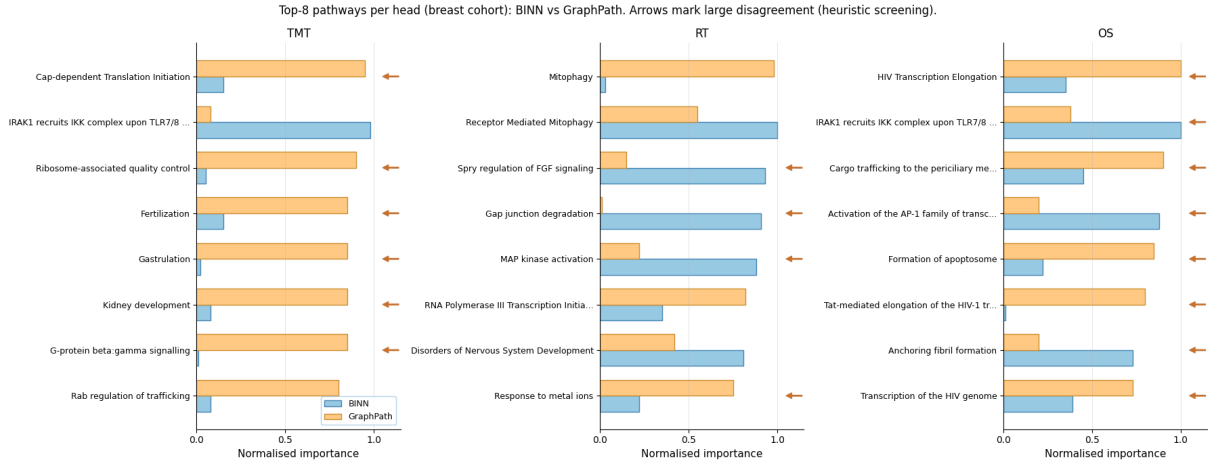

Figure 5: Normalised pathway importance scores for BINN and GraphPath across three clinical heads (breast cohort). Colours use the Okabe–Ito palette [30, 46]. Arrows flag pathways with a > 0.50 difference between models as a heuristic for manual review, not statistical significance (Section 5.4). Of 23 pathways, 7 fall outside PATH's ≥15-gene eligible set (Section 4). .

6 Discussion

Under a unified split with matched uncertainty quantification, model performance is task-dependent, but most inter-architecture differences are not statistically significant. Across treatment - assignment heads, no architecture significantly outperforms the others. The clearest exception is survival: BINN is the most consistent OS predictor and significantly outperforms both graph models on breast-cancer OS, suggesting that its multi-depth hierarchical supervision captures survival-related signals across pathway organization. We therefore avoid claiming that any model is best for TMT; although prostate is the most predictable cohort, all three architectures perform comparably. RT performance is generally weak and comparable across models, with paired-bootstrap tests (Table 4) rarely showing significant differences. This is consistent with, but does not prove, the hypothesis that radiation-treatment assignment depends largely on clinical factors not captured by pathway activity.

The pathway analysis highlights the interpretability of BINN and GraphPath. Both ranked the ERK/MAPK Targets pathway among the most influential across several cohorts and outcomes, suggesting a potentially meaningful biological signal [5]. PATH achieved strong predictive performance but lacks pathway-level interpretability, making its biological predictions harder to understand [4, 36].

Several limitations remain. The labels represent documented treatment exposure rather than treatment response, while OS is a short-term prognostic classification; validation on true response endpoints is therefore needed. Some cohorts have very few positive or negative cases, particularly thyroid TMT (n=7 positives) and prostate/thyroid OS, resulting in substantial uncertainty despite repeated stratified splits and paired-bootstrap tests. The study also uses repeated random splits rather than full k-fold cross-validation and lacks external validation. OS is treated as a fixed-horizon binary outcome and does not model censoring or time-to-event dynamics. Finally, PATH evaluates 1,431 of the 1,706 shared Reactome pathways because of its gene-membership filter, leaving 275 pathways with fewer than 15 genes. Of the 23 pathways identified in the BINN/GraphPath importance analysis (Fig. 5), 7 fall outside PATH's eligible set. Future work should standardize pathway selection across architectures, for example by applying a shared minimum gene threshold.

Overall, the strongest statistically supported finding is BINN's advantage for short-term survival (OS). For treatment assignment (TMT, RT), the architectures are largely comparable, while RT may benefit from integrating molecular and clinical information.

7 Conclusion

This study provides a unified benchmark for pathway-informed deep learning in cancer therapy and prognosis prediction. Using Reactome-derived pathway activity profiles from TCGA transcriptomic data, we evaluated BINN, GraphPath, and PATH across five solid-tumor cohorts and three clinically relevant outcomes: targeted molecular therapy (TMT), radiation therapy (RT), and overall survival (OS $\geq$ 6 months).

Pathway-guided architectures use biological knowledge for clinical prediction and give interpretable pathway-level readouts, but no single model beat the others across all cohorts and tasks: performance depends on both the clinical endpoint and the disease context. The choice of architecture should therefore follow the specific prediction objective, not an aggregate benchmark score.

The contribution here is less any single winner than the protocol: a shared data, preprocessing, training, and evaluation setup under which pathway-based models can be compared on equal terms, with confidence intervals and paired-bootstrap tests deciding which differences are real. Useful next steps are response-labelled cohorts, integration of clinical with molecular features (particularly for RT), and external validation on independent datasets. Before tools of this kind inform real treatment-assignment or prognosis decisions, clinicians need to know which architectural claims survive repeated evaluation rather than a single favourable split—which is what this benchmark is meant to establish.

References

[1] David A. Barbie, Pablo Tamayo, Jesse S. Boehm, So Young Kim, Susan E. Moody, Ian F. Dunn, Anna C. Schinzel, Peter Sandy, Etienne Meylan, Claudia Scholl, et al. 2009. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* 462, 7269 (2009), 108–112. doi:10.1038/nature08460

[2] Rajamanickam Baskar, Kuo Ann Lee, Richard Yeo, and Kheng-Wei Yeoh. 2012. Cancer and radiation therapy: current advances and future directions. *International Journal of Medical Sciences* 9, 3 (2012), 193–199. doi:10.7150/ijms.3635

[3] Harald Binder et al. 2020. Adaptive ERK Signalling Activation in Response to Therapy and In Silico Prognostic Evaluation of EGFR-MAPK in HNSCC. *British Journal of Cancer* (2020). doi:10.1038/s41416-020-0892-9

[4] Jim Clauwaert, Gerben Menschaert, and Willem Waegeman. 2021. Explainability in transformer models for functional genomics. *Briefings in Bioinformatics* 22, 5 (2021), bbab060. doi:10.1093/bib/bbab060

[5] Amardeep S. Dhillon, Suzanne Hagan, Oliver Rath, and Walter Kolch. 2007. MAP Kinase Signalling Pathways in Cancer. *Oncogene* 26, 22 (2007), 3279–3290. doi:10.1038/sj.onc.1210421

[6] Yifan Dou and Golrokh Mirzaei. 2025. MO-GCAN: multi-omics integration based on graph convolutional and attention networks. *Bioinformatics* 41, 8 (2025), btaf405. doi:10.1093/bioinformatics/btaf405

[7] Vijay Prakash Dwivedi and Xavier Bresson. 2021. A Generalization of Transformer Networks to Graphs. *AAAI Workshop on Deep Learning on Graphs: Methods and Applications* (2021). https://arxiv.org/abs/2012.09699

[8] Haitham A. Elmarakeby, Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H. AlDubayan, Keyan Salari, Steven Kregel, Camden Richter, et al. 2021. Biologically informed deep neural network for prostate cancer discovery. *Nature* 598, 7880 (2021), 348–352. doi:10.1038/s41586-021-03922-4

[9] Alexander S. Fisch et al. 2022. Precision Drugging of the MAPK Pathway in Head and Neck Cancer. *npj Genomic Medicine* (2022). doi:10.1038/s41525-022-00293-1

[10] Yan-Li Ge et al. 2013. MiR-200c Increases the Radiosensitivity of NSCLC Cell Line A549 by Targeting VEGF-VEGFR2 Pathway. *PLOS ONE* (2013). PMC3813610. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3813610/

[11] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, et al. 2022. The Reactome pathway knowledgebase 2022. *Nucleic Acids Research* 50, D1 (2022), D687–D692. doi:10.1093/nar/gkab1028

[12] Mary J. Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N. Brooks, et al. 2020. Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology* 38 (2020), 675–678. doi:10.1038/s41587-020-0546-8

[13] Erik Hartman, Aaron M. Scott, Christofer Karlsson, Tirthankar Mohanty, Suvi T. Vaara, Adam Linder, Lars Malmström, and Johan Malmström. 2023. Interpreting biologically informed neural networks for enhanced proteomic biomarker discovery and pathway analysis. *Nature Communications* 14 (2023), 5359. doi:10.1038/s41467-023-41146-4

[14] Koushik Howlader, Md Tauhidul Islam, and Wei Le. 2026. Graph Transformer-Based Pathway Embedding for Cancer Prognosis. arXiv:2604.16685 [cs.LG] https://arxiv.org/abs/2604.16685

[15] Zhi Huang, Xiaohui Zhan, Shuo Xiang, Travis S. Johnson, Bryan R. Helm, Christina Y. Yu, Jie Zhang, Paul Salama, Monther Rizkalla, Zhi Han, and Kun Huang. 2019. SALMON: Survival Analysis Learning With Multi-Omics Neural Networks on Breast Cancer. *Frontiers in Genetics* 10 (2019), 166. doi:10.3389/fgene.2019.00166

[16] Yuexu Jiang, Manish Sridhar Immadi, Duolin Wang, Shuai Zeng, Yen On Chan, Jing Zhou, Dong Xu, and Trupti Joshi. 2024. IRnet: Immunotherapy response

prediction using pathway knowledge-informed graph neural network. *Journal of Advanced Research* 72 (2024), 319–331. doi:10.1016/j.jare.2024.07.036

[17] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1412.6980

[18] Hiroki Kudo et al. 2025. Androgen Receptor Contributes to Radioresistance Through DNA Repair and Autophagy in AR-Positive Prostate Cancer Cells. bioRxiv. doi:10.64898/2025.12.03.690226

[19] Wei Lan, Haibo Liao, Qingfeng Chen, Ling-ling Zhu, Yi Pan, and Yi-Ping Phoebe Chen. 2024. DeepKEGG: a multi-omics data integration framework with biological insights for cancer recurrence prediction and biomarker discovery. *Briefings in Bioinformatics* 25, 3 (2024), bbae185. doi:10.1093/bib/bbae185

[20] Dohee Lee, Jaegyoon Ahn, and Jonghwan Choi. 2025. PathNetDRP: a novel biomarker discovery framework using pathway and protein–protein interaction networks for immune checkpoint inhibitor response prediction. *BMC Bioinformatics* 26, 1 (2025), 119. doi:10.1186/s12859-025-06125-0

[21] Min Li, M. Jin, Mingzhu Lou, Shaobo Deng, Lei Wang, and Hua Rao. 2025. Multiview-cooperated graph neural network enables novel multi-omics cancer subtype classification. *Computational Biology and Chemistry* (2025). doi:10.1016/j.compbiolchem.2025.108560

[22] Xiangmei Li, Bin Pan, Yao He, et al. 2025. PathHDNN: a pathway hierarchical-informed deep neural network framework for predicting immunotherapy response and mechanism interpretation. *Genome Medicine* 17 (2025), 152. doi:10.1186/s13073-025-01584-9

[23] Cheng-Pei Lin, Yann-Jen Ho, Yen-Peng Chiu, Yun Tang, You Sheng Paik, Guan Chen, Wei-Chih Huang, and Tzong-Yi Lee. 2026. MoAGNN: a multi-omics hierarchical graph neural network for subtype classification and prognosis prediction in lung adenocarcinoma. *Briefings in Bioinformatics* 27, 1 (2026), bbaf735. doi:10.1093/bib/bbaf735

[24] Xiaofan Liu, Yuhuan Tao, Zilin Cai, Pengfei Bao, Hongli Ma, Kexing Li, Mingyue Li, Yongchang Zhu, and Zhi John Lu. 2024. Pathformer: a biological pathway informed transformer for disease diagnosis and prognosis using multi-omics data. *Bioinformatics* 40, 5 (2024), btae316. doi:10.1093/bioinformatics/btae316

[25] Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1711.05101

[26] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30. https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

[27] Teng Ma and Jianxin Wang. 2024. GraphPath: a graph attention model for molecular stratification with interpretability based on the pathway–pathway interaction network. *Bioinformatics* 40, 4 (2024), btae165. doi:10.1093/bioinformatics/btae165

[28] Teng Ma, Haochen Zhao, Qichang Zhao, and Jianxin Wang. 2024. Cox-Path: Biological Pathway-Informed Graph Neural Network for Cancer Survival Prediction. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '24)*. 70:1–70:6. doi:10.1145/3698587.3701397

[29] Hye-Young Min and Ho-Young Lee. 2022. Molecular targeted therapy for anticancer treatment. *Experimental & Molecular Medicine* 54, 10 (2022), 1670–1694. doi:10.1038/s12276-022-00864-3

[30] Masataka Okabe and Kei Ito. 2008. Color universal design (CUD): How to make figures and presentations that are friendly to colorblind people. *Color Universal Design Organization technical note* (2008). https://jfly.uni-koeln.de/color/

[31] Ravi B. Parikh, Christopher Manz, Corey Chivers, et al. 2019. Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA Network Open* 2, 10 (2019), e1915997. doi:10.1001/jamanetworkopen.2019.15997

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems* 32 (2019). https://papers.nips.cc/paper_files/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html

[34] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. doi:10.1609/aaai.v32i1.11671

[35] Olivier B Poirion, Zheng Jing, Kuldeep Chaudhary, Sijia Huang, and Lana X Garmire. 2021. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. *Genome Medicine* 13, 1 (2021), 112. doi:10.1186/s13073-021-00930-x

[36] Rikard Rosenbacke, Åsa Melhus, Martin McKee, and David Stuckler. 2024. How Explainable Artificial Intelligence Can Increase or Decrease Clinicians' Trust in AI Applications in Health Care: Systematic Review. *JMIR AI* 3 (2024), e53207. doi:10.2196/53207

[37] Chris J. Sidey-Gibbons, Charlotte Sun, Alina Schneider, et al. 2022. Predicting 180-day mortality for women with ovarian cancer using machine learning and patient-reported outcome data. *Scientific Reports* 12, 1 (2022), 20614. doi:10.1038/s41598-022-22614-1

[38] Srinivasa Prasad Sisinthy and Dhruv L. Bhatt. 2021. Everything Old Is New Again: Drug Repurposing Approach for NSCLC Targeting MAPK Signaling Pathway. *Frontiers in Oncology* 11 (2021). doi:10.3389/fonc.2021.741326

[39] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102, 43 (2005), 15545–15550. doi:10.1073/pnas.0506580102

[40] The Cancer Genome Atlas Network. 2012. Comprehensive molecular portraits of human breast tumours. *Nature* 490, 7418 (2012), 61–70. doi:10.1038/nature11412

[41] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1710.10903

[42] Di Wang, Chupei Tang, Junxiao Kong, Jixiu Zhai, Moyu Tang, and Tianchi Lu. 2026. PathMoG: A Pathway-Centric Modular Graph Neural Network for Multi-Omics Survival Prediction. arXiv:2604.24371 [cs.LG] https://arxiv.org/abs/2604.24371

[43] Asim Waqas, Aakash Tripathi, Sabeen Ahmed, Ashwin Mukund, Hamza Farooq, Joseph O. Johnson, Paul A. Stewart, Mia Naeini, Matthew B. Schabath, and Ghulam Rasool. 2025. Self-Normalizing Multi-Omics Neural Network for Pan-Cancer Prognostication. *International Journal of Molecular Sciences* 26, 15 (2025), 7358. doi:10.3390/ijms26157358

[44] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* 45, 10 (2013), 1113–1120. doi:10.1038/ng.2764

[45] Gang Wen and Limin Li. 2023. FGCNSurv: dually fused graph convolutional network for multi-omics survival prediction. *Bioinformatics* 39, 8 (2023), btad472. doi:10.1093/bioinformatics/btad472

[46] Bang Wong. 2011. Points of view: Color blindness. *Nature Methods* 8, 6 (2011), 441. doi:10.1038/nmeth.1618

[47] Magdalena Wysocka, Oskar Wysocki, Marie Zufferey, Dónal Landers, and André Freitas. 2023. A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data. *BMC Bioinformatics* 24 (2023), 198. doi:10.1186/s12859-023-05262-8

[48] Hongxi Yan, Dawei Weng, Dongguo Li, Yu Gu, Wenjie Ma, and Qingjie Liu. 2024. Prior knowledge-guided multilevel graph neural network for tumor risk prediction and interpretation via multi-omics data integration. *Briefings in Bioinformatics* 25, 3 (2024), bbae184. doi:10.1093/bib/bbae184

[49] Jiayang Zhang, Yilin Che, Rongrong Liu, Zhicheng Wang, and Weiwu Liu. 2025. Deep learning–driven multi-omics analysis: enhancing cancer diagnostics and therapeutics. *Briefings in Bioinformatics* 26, 4 (2025), bbaf440. doi:10.1093/bib/bbaf440

[50] Tinghe Zhang, Md Hasib, Yu-Chiao Chiu, Zhizhong Han, Yu-Fang Jin, Mario Flores, Yidong Chen, and Yufei Huang. 2022. Transformer for Gene Expression Modeling (T-GEM): An Interpretable Deep Learning Model for Gene Expression-Based Phenotype Predictions. *Cancers* 14, 19 (2022), 4763. doi:10.3390/cancers14194763

[51] Lianhe Zhao, Qiongye Dong, Chunlong Luo, Yang Wu, Dechao Bu, Xiaoning Qi, Yufan Luo, and Yi Zhao. 2021. DeepOmix: A scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis. *Computational and Structural Biotechnology Journal* 19 (2021), 2719–2725. doi:10.1016/j.csbj.2021.04.067

[52] Yue Zhao et al. 2015. Prognostic Values of ERK1/2 and p-ERK1/2 Expressions for Poor Survival in Non-Small Cell Lung Cancer. *Tumor Biology* (2015). doi:10.1007/s13277-015-3048-4

[53] Rong Zheng et al. 2015. TAT-ODD-p53 Enhances Radiosensitivity of Hypoxic Breast Cancer Cells by Inhibiting Parkin-Mediated Mitophagy. *Oncotarget* (2015). PMC4627318. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4627318/

[54] Jiening Zhu, John H. Oh, Anish K. Simhal, Rena Elkin, Larry Norton, Joseph O. Deasy, and Allen Tannenbaum. 2023. Geometric graph neural networks on multi-omics data to predict cancer survival outcomes. *Computers in Biology and Medicine* 163 (2023), 107117. doi:10.1016/j.compbiomed.2023.107117

[55] Payam Zohari and Mostafa Haghir Chehreghani. 2025. Graph Neural Networks in Multi-Omics Cancer Research: A Structured Survey. *arXiv* (2025). arXiv:2506.17234 https://arxiv.org/abs/2506.17234

A Architecture and Metric Details

A.1 Performance Metrics

We evaluate model performance using five complementary metrics across all benchmark tasks.

AUROC measures the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative:

$$\text{AUROC} = \int_0^1 \text{TPR}\big(\text{FPR}^{-1}(t)\big)\, dt \tag{1}$$

AUPRC summarizes the precision-recall trade-off and is particularly informative under class imbalance, which is common in cancer genomics datasets:

$$\text{AUPRC} = \sum_k (R_k - R_{k-1})\, P_k \tag{2}$$

F1 Score is the harmonic mean of precision (P) and recall (R) for the positive class of each clinical head, evaluated at the validation-tuned decision threshold:

$$\text{F1} = \frac{2\,\text{P}\cdot\text{R}}{\text{P}+\text{R}}, \qquad \text{P} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \qquad \text{R} = \frac{\text{TP}}{\text{TP}+\text{FN}}. \tag{3}$$

Because each head is a single binary decision, we report the positive-class (binary) F1 per head rather than a support-weighted average across classes; this matches the quantity computed by our evaluation code.

Accuracy measures the fraction of correctly classified samples:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{y}_i = y_i] \tag{4}$$

Confusion Matrix provides a full breakdown of predictions across classes, where entry C_{ij} denotes samples of true class i predicted as class j:

$$C = \begin{pmatrix} \text{TP} & \text{FN} \\ \text{FP} & \text{TN} \end{pmatrix} \tag{5}$$

A.2 BINN - Sparse Reactome Hierarchy

Architecture Overview: BINN integrates the Reactome parent-child hierarchy directly into the network by imposing sparse masked-linear connections between layers, together with auxiliary classification heads at each depth. This design ensures that gradient signals reach every intermediate node during training, making gradient-×-input attribution across the hierarchy well-posed [26].

Pathway Hierarchy: Starting from 1,706 Reactome-matched input pathways, the model traverses parent-child edges upward through three hidden layers, yielding a four-layer network with dimensions:

$$1{,}706 \to 656 \to 306 \to 163. \tag{6}$$

Pathways that terminate before the maximum depth are copied forward, ensuring every pathway is represented at every layer.

Sparse Masked Weights: Between consecutive layers ℓ and $\ell+1$, the Reactome hierarchy defines a binary connectivity mask $M^{(\ell)} \in \{0,1\}^{|L_{\ell+1}|\times|L_\ell|}$, where:

$$M_{ij}^{(\ell)} = 1 \iff \text{node } j \text{ in layer } \ell \text{ is a Reactome child of node } i \text{ in layer } \ell{+}1, \tag{7}$$

or node j is a terminal node copied forward to position i. Weights are initialised with Kaiming uniform initialisation and zeroed outside the mask prior to the first forward pass. Sparsity is enforced throughout training via the elementwise product $W^{(\ell)} \odot M^{(\ell)}$.

Forward Pass: Each hidden block applies the masked linear transformation followed by Tanh activation, Batch Normalisation, and Dropout ($p = 0.2$):

$$\begin{aligned} z^{(\ell)} &= \big(W^{(\ell)} \odot M^{(\ell)}\big)\, h^{(\ell)} + b^{(\ell)}, \\ h^{(\ell+1)} &= \text{Dropout}\Big(\text{BN}\Big(\tanh\Big(z^{(\ell)}\Big)\Big)\Big). \end{aligned} \tag{8}$$

Multi-Layer Supervision: An auxiliary classifier head Linear($|L_\ell| \to 3$) is attached after every layer, including the raw input ($\ell = 0$), producing three logits per layer. The final predicted probability for head h is the mean of all four sigmoid outputs:

$$\hat{p}_h(x) = \frac{1}{4}\sum_{\ell=0}^{3} \sigma\Big(H^{(\ell)}\Big(h^{(\ell)}\Big)\,[h]\Big), \tag{9}$$

where $H^{(\ell)}$ denotes the classifier at layer ℓ. Attaching a loss signal at every depth ensures that each intermediate node receives direct supervision, a prerequisite for meaningful attribution across the hierarchy.

A.3 GraphPath - Multi-Head Graph Attention

Architecture Overview: GraphPath treats the 1,706 Reactome pathways as nodes in a graph and updates their embeddings through a multi-head graph attention network (GAT). Because pathways interact and co-regulate one another rather than acting as independent units, each pathway aggregates information from its biologically annotated neighbours before the classification step.

Pathway Adjacency Graph: A symmetric binary adjacency matrix $A \in \{0,1\}^{N\times N}$ is constructed such that:

$$A_{pq} = 1 \iff \text{pathways } p \text{ and } q \text{ share a Reactome parent-child or sibling relationship.} \tag{10}$$

This yields 4,672 undirected edges with a mean node degree of 5.5. Self-loops are added so that each node attends to its own embedding during message passing.

Input Projection: Each scalar ssGSEA score x_p is projected to a d-dimensional embedding ($d = 64$) via a shared weight matrix $W_{\text{proj}} \in \mathbb{R}^{1\times d}$ and a pathway-specific bias $b_p \in \mathbb{R}^d$, followed by Tanh activation:

$$h_p^{(0)} = \tanh\big(x_p \cdot W_{\text{proj}} + b_p\big). \tag{11}$$

The pathway-specific bias captures baseline differences in pathway activity, while the shared projection weight preserves parameter efficiency.

Multi-Head Graph Attention: Node embeddings are updated by a single GAT layer with $K = 3$ attention heads and ELU output activation. For each head k, the raw attention score and normalised

coefficient are:

$$e_{ij}^{(k)} = a^{(k)\top}\left[W^{(k)}h_i \,\|\, W^{(k)}h_j\right],$$
$$\alpha_{ij}^{(k)} = \mathrm{softmax}_{j\in\mathcal{N}(i)}\,\mathrm{LeakyReLU}_{0.2}\left(e_{ij}^{(k)}\right). \tag{12}$$

Non-edges ($A_{ij} = 0$) are masked to $-\infty$ before the softmax, confining attention to biologically annotated neighbours. Attention dropout ($p = 0.4$) is applied during training. Per-head embeddings are concatenated to yield a node representation of dimension $K \cdot d$.

Readout and Classification: A shared Linear($Kd \to 1$) layer followed by Tanh collapses each node to a scalar readout. The resulting N-dimensional vector is passed through a fully-connected layer mapping to three binary head logits, followed by sigmoid activation.

A.4 PATH - Edge-Aware Graph Transformer

Architecture Overview: PATH constructs a weighted pathway graph from gene-set overlap and augments a Graph Transformer with Laplacian positional encodings and learnable edge-conditioned attention biases. This allows the model to represent graded biological similarity between pathways and to partially rewire the prior graph when the data warrant it.

Weighted Pathway Adjacency: Edge weights are defined as the Jaccard similarity of Reactome gene memberships:

$$A_{pq} = \frac{|G_p \cap G_q|}{|G_p \cup G_q|}, \quad |G_p| \geq 15, \tag{13}$$

computed from the Reactome GMT file and renormalised to $[0, 1]$. Only pathways with at least 15 annotated genes qualify as nodes, yielding 1,431 nodes, 243,210 weighted edges, and a mean node degree of 339.9. This filtering criterion is applied directly to the same 1,706-pathway Reactome set used by BINN and GraphPath (Section 4); PATH's 1,431-pathway node set is therefore a *strict subset* of the shared input space, not an independently defined one. The 275 excluded pathways (16.1% of 1,706) are removed solely because they fall below the $|G_p| \geq 15$ gene-membership threshold required to compute a well-defined Jaccard edge weight—no pathway outside the original 1,706 is introduced. Consequently, any comparison between PATH and the other two architectures is bounded by a proper-subset relationship rather than an open-ended mismatch: a pathway appearing in the BINN/GraphPath importance ranking (Fig. 5) but absent from PATH's node set is, by construction, excluded from PATH's attention-derived analysis rather than mis-ranked by it. The dense connectivity reflects the substantial gene-set overlap characteristic of the Reactome hierarchy and motivates a soft structural mask rather than hard adjacency truncation.

Input Projection and Positional Encoding: To ensure a fair comparison, we replaced PATH's original Stage 1 (FiLM-modulated gene embeddings) and Stage 2 (intra-pathway attention pooling) with the same per-pathway scalar projection used in GraphPath [34]. This setup uses a shared weight and a pathway-specific bias to project each score into a 64-dimensional space ($\mathbb{R}^d$, $d = 64$), followed by a Tanh activation.

Laplacian positional encodings are added to endow the model with structural awareness. The top-$k = 16$ non-trivial eigenvectors of the symmetrically normalised Laplacian $L = I - D^{-1/2}AD^{-1/2}$ are concatenated with the normalised node degree $\deg(p)/(N + \epsilon)$, forming a $(k{+}1)$-dimensional positional feature per node. These features are projected to $\mathbb{R}^d$ via a learnable linear layer and added to node embeddings before the transformer stack. Eigenvector signs are randomly flipped per epoch during training to resolve sign ambiguity.

Edge-Aware Transformer Blocks. PATH employs $L = 2$ transformer blocks, each computing scaled dot-product multi-head self-attention ($H = 4$ heads) augmented by two structural biases.

Soft structural mask. Non-edges are heavily but softly down-weighted:

$$m_{pq}^{(\mathrm{struct})} = \begin{cases} 0 & \text{if } A_{pq} > 0 \text{ or } p = q, \\ -10 & \text{otherwise.} \end{cases} \tag{14}$$

Keeping non-edges gradient-reachable allows the model to rewire the prior graph when the data provide sufficient evidence.

Edge-conditioned attention bias. A learnable per-layer scalar edge feature is aggregated over attention heads to produce a continuous bias on the attention logits:

$$\phi_{pq}^{(\ell)} = \frac{1}{H}\sum_{h=1}^{H} \log\,\mathrm{softplus}\left(w_h^{(\ell)}\, e_{pq}^{(\ell)} + b_h^{(\ell)}\right) + \epsilon, \tag{15}$$

where $e_{pq}^{(\ell)}$ is initialised from the Jaccard weight. The combined bias $m_{pq}^{(\mathrm{struct})} + \phi_{pq}^{(\ell)}$ is added to the raw attention logits before the softmax, grounding attention in biological pathway similarity while permitting data-driven refinement.

Following attention, a residual-wrapped GELU feed-forward network (expansion factor 4) with per-token Batch Normalisation updates node features. A parallel two-layer MLP with Batch Normalisation updates edge features between transformer blocks.

Readout and Classification. Node tokens are collapsed to a graph-level embedding via attention-weighted readout:

$$g = \sum_p w_p\, x_p^{(L)},$$
$$w_p = \mathrm{softmax}_p\left(v^\top \tanh\left(U x_p^{(L)}\right)\right), \tag{16}$$

where $U \in \mathbb{R}^{d\times(d/2)}$ and $v \in \mathbb{R}^{d/2}$ are learnable parameters. This soft attention over nodes identifies the pathways most informative for the three prediction targets. The graph embedding g is then passed through a classification head:

$$\mathrm{Linear}(d \to d) \to \mathrm{BN} \to \mathrm{GELU} \to \mathrm{Dropout}(p = 0.2) \to \mathrm{Linear}(d \to 3), \tag{17}$$

producing three binary head logits.

A.5 Loss, Training, and Reproducibility

Objective Function: All three models minimise a per-head class-weighted binary cross-entropy loss. Positive-class weights are computed on the training fold as $w_h = \mathrm{neg}_h/\mathrm{pos}_h$ and clipped to $[0.1, 20]$ to prevent extreme imbalance from destabilising training. For BINN, the loss is applied to the averaged sigmoid outputs; for GraphPath and PATH,

it is applied to raw logits via `BCEWithLogitsLoss` (or `F.binary_cross_entropy_with_logits`) for numerical stability.

Optimisers and Learning Rate Schedules: Optimiser choices follow each reference implementation:

Table 3: Optimiser configuration per model.

	BINN	GraphPath	PATH
Optimiser	Adam [17]	SGD (momentum 0.9)	AdamW [25]
LR	10^{-3}	5×10^{-2}	10^{-4}
Weight Decay	10^{-3}	5×10^{-2}	5×10^{-4}
Batch Size	32	32	16

All three models employ ReduceLROnPlateau scheduling on validation loss with a reduction factor of 0.5 and plateau patience of 7, 8, and 10 epochs for BINN, GraphPath, and PATH respectively.

Early Stopping: Training is terminated when validation loss fails to improve for 20 consecutive epochs (BINN) or 25 consecutive epochs (GraphPath, PATH). For PATH, a minimum of 25 training epochs is required before the patience counter activates, preventing premature termination during the characteristically slow initial convergence of the transformer stack. The checkpoint with the lowest validation loss is restored prior to evaluation.

B Full Paired-Bootstrap Significance Tests

Table 4: Paired-bootstrap significance tests on test AUROC (10^4 resamples of the identical held-out folds). ΔAUROC is (model A − model B); a 95% CI excluding 0 (equivalently $p < 0.05$, shown in bold) indicates a significant difference. † marks underpowered cells with fewer than 15 pooled minority-class test samples (thyroid and the near-universal-prevalence OS cohorts), whose bootstrap estimates are unstable and are not interpreted as significant.

Cohort	Head	Contrast (A vs B)	ΔAUROC [95% CI]	p
Breast	TMT	PATH vs BINN	-0.038 [-0.174, +0.105]	0.578
Breast	TMT	PATH vs GraphPath	-0.009 [-0.150, +0.133]	0.924
Breast	TMT	GraphPath vs BINN	-0.029 [-0.181, +0.129]	0.707
Breast	RT	PATH vs BINN	-0.006 [-0.078, +0.068]	0.881
Breast	RT	PATH vs GraphPath	-0.006 [-0.074, +0.064]	0.865
Breast	RT	GraphPath vs BINN	+0.000 [-0.071, +0.073]	1.000
Breast	OS	PATH vs BINN	**-0.140 [-0.226, -0.053]**	0.001
Breast	OS	PATH vs GraphPath	-0.040 [-0.133, +0.055]	0.410
Breast	OS	GraphPath vs BINN	**-0.101 [-0.178, -0.026]**	0.010
Lung	TMT	PATH vs BINN	-0.028 [-0.096, +0.042]	0.431
Lung	TMT	PATH vs GraphPath	+0.029 [-0.043, +0.101]	0.437
Lung	TMT	GraphPath vs BINN	-0.057 [-0.139, +0.028]	0.184
Lung	RT	PATH vs BINN	-0.014 [-0.115, +0.092]	0.793
Lung	RT	PATH vs GraphPath	+0.055 [-0.041, +0.147]	0.246
Lung	RT	GraphPath vs BINN	-0.069 [-0.175, +0.038]	0.209
Lung	OS	PATH vs BINN	-0.034 [-0.140, +0.073]	0.513
Lung	OS	PATH vs GraphPath	+0.025 [-0.105, +0.158]	0.697
Lung	OS	GraphPath vs BINN	-0.060 [-0.176, +0.060]	0.330
Prostate	TMT	PATH vs BINN	-0.035 [-0.094, +0.018]	0.212
Prostate	TMT	PATH vs GraphPath	-0.002 [-0.064, +0.060]	0.945
Prostate	TMT	GraphPath vs BINN	-0.033 [-0.076, +0.008]	0.114
Prostate	RT	PATH vs BINN	-0.030 [-0.083, +0.024]	0.271
Prostate	RT	PATH vs GraphPath	+0.008 [-0.074, +0.090]	0.861
Prostate	RT	GraphPath vs BINN	-0.038 [-0.105, +0.030]	0.266
Prostate	OS	PATH vs BINN	-0.322 [-0.546, -0.096]	0.008†
Prostate	OS	PATH vs GraphPath	-0.195 [-0.452, +0.050]	0.125†
Prostate	OS	GraphPath vs BINN	-0.127 [-0.292, +0.042]	0.142†
Head & Neck	TMT	PATH vs BINN	-0.052 [-0.148, +0.044]	0.283
Head & Neck	TMT	PATH vs GraphPath	-0.011 [-0.113, +0.087]	0.832
Head & Neck	TMT	GraphPath vs BINN	-0.041 [-0.103, +0.020]	0.196
Head & Neck	RT	PATH vs BINN	-0.069 [-0.173, +0.032]	0.189
Head & Neck	RT	PATH vs GraphPath	-0.010 [-0.111, +0.091]	0.846
Head & Neck	RT	GraphPath vs BINN	**-0.059 [-0.106, -0.013]**	0.013
Head & Neck	OS	PATH vs BINN	-0.163 [-0.367, +0.030]	0.098†
Head & Neck	OS	PATH vs GraphPath	-0.163 [-0.328, -0.008]	0.038†
Head & Neck	OS	GraphPath vs BINN	-0.001 [-0.105, +0.092]	0.975†
Thyroid	TMT	PATH vs BINN	+0.120 [-0.091, +0.367]	0.291†
Thyroid	TMT	PATH vs GraphPath	+0.184 [-0.167, +0.569]	0.320†
Thyroid	TMT	GraphPath vs BINN	-0.064 [-0.425, +0.213]	0.760†
Thyroid	RT	PATH vs BINN	+0.070 [-0.112, +0.253]	0.461
Thyroid	RT	PATH vs GraphPath	+0.032 [-0.144, +0.214]	0.720
Thyroid	RT	GraphPath vs BINN	+0.038 [-0.076, +0.152]	0.526
Thyroid	OS	PATH vs BINN	+0.593 [+0.462, +0.722]	<0.001†
Thyroid	OS	PATH vs GraphPath	+0.704 [+0.574, +0.815]	<0.001†
Thyroid	OS	GraphPath vs BINN	-0.111 [-0.208, -0.019]	0.034†

C Training and Reproducibility

Reproducibility and availability: All models are implemented in PyTorch 2.0 [32], with scikit-learn [33] used for metric computation. Random seeds are fixed identically across Python, NumPy, and PyTorch for all models and cohorts; the repeated-split sweep is driven by `slurm/20_cv.sbatch` and aggregated by `paper/scripts/aggregate_cv.py`. The single-seed figures are regenerated by executing `scripts/run_all.sh` within the `binn/`, `graphpath/`, and `path/` directories, followed by `paper/build.sh`. Source code, the exact per-cohort, per-seed split index files, and the per-sample test predictions used for the bootstrap tests are released so that every reported number and split can be reproduced exactly.