Research Article

COX-GAT: A Leakage-Controlled Graph Attention Network For Pathway Dysregulation Scoring And Multi-Omics Prognostic Biomarker Discovery in Clear Cell Renal Cell Carcinoma

Jiyad Mohammed KV[1,*], Chandramani Chaudhary [2]

[1]Research Scholar, Department of Computer Science and Engineering, National Institute of Technology Calicut, Kozhikode, Kerala, India.
[2]Assistant Professor, Department of Computer Science and Engineering, National Institute of Technology Calicut, Kozhikode, Kerala, India.

*Corresponding author: Jiyad Mohammed KV | Email: jiyad99@gmail.com

Abstract

Clear cell renal cell carcinoma (ccRCC) shows marked prognostic heterogeneity beyond conventional staging, yet many multi-omics graph neural network studies risk information leakage and few are validated on independent cohorts. We introduce Cox-GAT, a leakage-audited graph attention framework that integrates transcriptomic, copy-number and clinical data on a protein–protein interaction graph and is trained under a Cox partial-likelihood objective. Edge weights, feature normalization and risk thresholds were derived from training data only and verified by automated checks; the pooled Kaplan–Meier stratification and the pathway ranking are exceptions, using the pooled out-of-fold median risk score. Across 15 patient-level cross-validation folds on TCGA-KIRC, Cox-GAT achieved a mean concordance index of 0.7697 with significant stratification in every fold. Pathway scoring of learned embeddings identified 31 prognostic Reactome pathways, led by cholesterol biosynthesis, SUMOylation and the TCA cycle, and gradient attribution nominated candidate prognostic genes including ANGPTL4, CYP51A1, ALDOB and AXIN2. Six of the ten top pathways and three of the four highlighted genes were consistent with prior renal cancer literature. Without refitting, the models generalized to an independent CPTAC-3 cohort (ensemble C-index 0.707). Under identical folds, Cox-GAT showed a modest advantage over DeepSurv, GCN-surv and GraphSAGE-surv. The largest ablation losses followed removal of the clinical covariates or the residual connection and random rewiring of the PPI graph, whereas the real graph exceeded no graph by only 0.0065 (untested). Gene selection used the full cohort and attribution a single fold. Cox-GAT offers a reproducible, externally validated multi-omics survival framework for ccRCC.

Keywords: Clear cell renal cell carcinoma; graph attention network; data leakage; pathway dysregulation; external validation

Introduction

Clear cell renal cell carcinoma (ccRCC) accounts for 75–80% of renal cell carcinoma diagnoses, and survival falls sharply once the disease has metastasized[1]. Patients who share the same tumor–node–metastasis (TNM) stage and grade follow markedly different clinical courses, so anatomical staging alone cannot identify who will relapse or die[1]. Inactivation of the von Hippel–Lindau (VHL) tumor suppressor activates hypoxia-inducible factor (HIF) signaling and drives metabolic reprogramming, including down-regulation of tricarboxylic acid (TCA) cycle genes in aggressive tumors[2], and broader genomic instability further diversifies the transcriptome[2]. Even so, risk stratification still relies chiefly on TNM stage and histological grade[1].

Because ccRCC pathways are highly interconnected, biomarkers studied in isolation rarely capture the systems-level dysregulation that determines outcome. Public resources such as The Cancer Genome Atlas pair transcriptomic and copy-number profiles with long-term survival[3], and graph attention networks[4], which learn how strongly each neighbor should contribute, suit the encoding of genes as nodes of protein–protein interaction (PPI) networks, as in prior knowledge-guided survival models[5,6]. Two weaknesses limit such models. First, many pipelines estimate normalization statistics and network edges on the entire cohort before partitioning it, which inflates reported performance[7]. Second, few have been evaluated on a genuinely independent cohort, although CPTAC-3 now makes this feasible[8].

Here we introduce Cox-GAT, an edge-weighted graph attention network trained under a Cox partial-likelihood objective on TCGA-KIRC RNA-seq, copy-number and clinical data. The study makes four contributions. First, every fold-dependent quantity, including PPI edge weights, feature normalization and risk-group thresholds, is derived from training data only and verified by automated audits, with the exceptions disclosed under Cross-validation design and leakage controls. Second, learned gene embeddings are converted into a pathway dysregulation score that links the survival model to Reactome biology. Third, gradient attribution nominates candidate prognostic genes within the associated pathways. Fourth, all fold models are applied without refitting to an independent CPTAC-3 cohort and are benchmarked against re-implemented baselines and single-component ablations under an identical protocol.

Materials and Methods

The end-to-end workflow, including the automated leakage audits applied at each stage, is summarized in Fig. 1.

Cohort and data acquisition

RNA-seq (STAR gene counts), somatic copy-number and clinical data for TCGA-KIRC were obtained from the Genomic Data Commons (GDC)[9] and cBioPortal[10]. STAR-count files for 582 samples were aggregated into a raw counts matrix, and Ensembl identifiers were mapped to approved HGNC symbols[11], retaining 41,924 of 60,616 genes. Sample type was assigned from the fourth field of the TCGA barcode. Copy-number data were GISTIC2.0-thresholded discrete calls (−2 to 2)[12], with normal samples assigned CNV = 0. Clinical variables (vital status, overall survival time, age, AJCC stage and histologic grade) were parsed from cBioPortal tables. The raw cohort comprised 500 primary tumor and 69 adjacent-normal samples (569 in total). Only public, de-identified TCGA and CPTAC-3 data were used; no ethics approval or consent was required.

Differential expression analysis

Tumor-versus-normal differential expression was computed on all 569 raw samples (500 tumor, 69 normal), before CNV alignment, with PyDESeq2 v0.5.4[13], a Python implementation of the DESeq2 negative-binomial framework[14]. Low-count genes were pre-filtered (41,924 to 25,648 genes), and the design formula was ~ tissue_type (normal as reference). Genes with a Benjamini–Hochberg-adjusted P value (padj)[15] below 0.05 and an absolute log2 fold change above 1 were defined as differentially expressed genes (DEGs), yielding 9,997 genes (7,335 up-regulated and 2,662 down-regulated in tumor). Only the tissue-type label, never survival outcome or fold membership, was used. Because differential expression was computed once on the full cohort rather than within each training fold, held-out patients contributed to DEG selection; this is disclosed as a limitation (see Discussion).

Sample alignment, variance stabilization and copy-number processing

RNA and CNV profiles were aligned at the sample level, retaining 496 of the 500 tumor samples and all 69 normal samples (565 in total). The 496 tumor samples represented 491 unique patients, who form the modeling cohort. Variance-stabilizing transformation (VST)[14] was applied to the 565 aligned samples and restricted to DEG rows. It was run blind to sample groups, using neither tissue-type labels nor survival or fold information, and is therefore label-free; it was, however, fitted once on all 565 samples, including normal samples and held-out patients, and is therefore not fold-isolated.

GISTIC2 calls[12] were restricted to DEG symbols: 5,688 of 9,997 DEGs (56.9%) had a matching CNV probe, and the remainder were assigned CNV = 0. CNV values were left uncentered at this stage; centering on the training-tumor mean is performed within each fold (see Feature tensor assembly) so that no validation or test statistics enter the normalization.

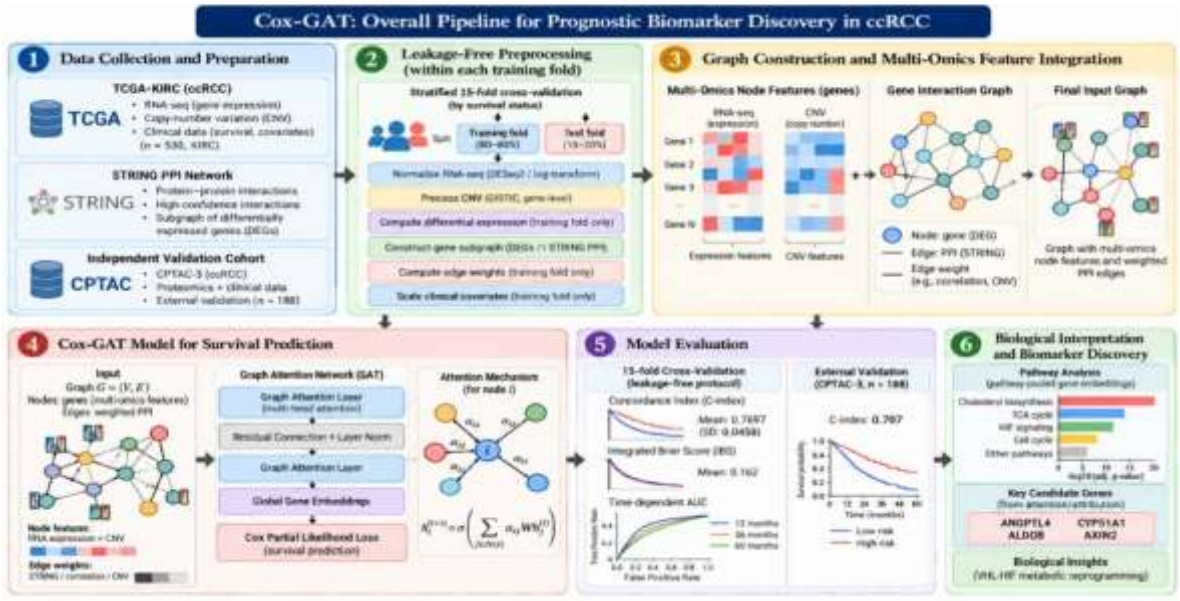


Fig. 1 — Overview of the leakage-controlled Cox-GAT pipeline. TCGA-KIRC RNA-seq, CNV and clinical data are mapped onto a STRING protein–protein interaction (PPI) subgraph of differentially expressed genes. Normalization, edge weights and covariate scaling are fitted on training partitions only within 15 cross-validation folds. A graph attention network trained with the Cox partial-likelihood loss is evaluated internally and on the independent CPTAC-3 cohort, followed by pathway scoring and gradient-based gene attribution.

Cross-validation design and leakage controls

All models were developed with repeated, stratified, patient-level five-fold cross-validation using three random seeds (42, 123, 456), giving 15 train/validation/test partitions in which no patient appears in more than one partition. Normal samples were used only for differential expression and VST. Stratification used a combined event × stage-bin label (early, stage I–II; late, stage III–IV), and each fold's training-plus-validation set was further divided 90%/10% into training and validation patients. Across the 491 eligible patients, folds contained on average 353 training, 40 validation and 98 test patients.

Four automated checks passed before modeling: (i) no patient overlap among training, validation and test sets within any fold; (ii) each patient in the test set exactly once per seed; (iii) a test-fold event-rate standard deviation of 0.002 across folds, far below the pre-specified tolerance of 0.08; and (iv) coverage of all 491 patients. Every subsequent preprocessing statistic (PPI edge weights, feature normalization and risk-group thresholds) was derived exclusively from each fold's training-tumor partition and applied unchanged to its validation and test patients, with mean-shift audits at each stage. Two analyses are exceptions: the pooled Kaplan–Meier stratification and the pathway ranking (see Pathway dysregulation scoring) use the median of the aggregated out-of-fold risk score across all 491 patients. The cohort-level differential expression and blind VST precede the split and use no survival information but are not fold-isolated.

Graph construction

PPI subgraph. DEGs were mapped onto the STRING v12 (Homo sapiens) interaction network[16], retaining undirected, deduplicated edges with a combined score of at least 600 between two DEG-encoded proteins. This yielded 4,414 DEG nodes (44.2% of the 9,997 DEGs) and 37,850 base edges; the remaining DEGs lacked a qualifying partner and were excluded (see Discussion).

Fold-specific edge weights. Pearson correlations between DEG expression profiles were recomputed for each of the 15 folds using that fold's training-tumor VST values only, and edges with $|r| < 0.3$ were pruned. This produced 15 distinct edge sets, each used solely by its own fold model.

Reactome pathway mapping. Gene–pathway membership was taken from Ensembl2Reactome[17], restricted to Homo sapiens, mapped to HGNC symbols and filtered to DEGs. Pathways with 5–500 DEG members were retained, and redundancy was reduced by a greedy Jaccard filter (pathways sorted from largest to smallest; a pathway was kept only if its Jaccard overlap with every kept pathway was below 0.7), yielding 807 non-redundant pathways (median 12 DEG members; maximum 469).

Feature tensor assembly

For each fold, node features comprised RNA expression z-scored with the training-tumor mean and standard deviation, GISTIC2 CNV centered on the training-tumor mean, and node degree in the fold's pruned PPI graph, min–max scaled to [0, 1]. Three clinical covariates were added per patient: age (z-scored on training-tumor statistics), stage (ordinal, divided by 4) and grade (same scheme, imputed to 0.5 when missing). All statistics were fitted on training-tumor samples only, as verified directly: training-set RNA z-scores had a mean of approximately 0.00003 by construction, whereas test-set z-scores retained a mean of approximately 0.015. Eleven automated shape, range, NaN and leakage checks passed for every fold.

Cox-GAT architecture

Cox-GAT is a two-layer, edge-weighted graph attention network[4] with a residual connection and a clinically augmented Cox risk head, implemented with PyTorch Geometric[18]. Each of the 4,414 PPI nodes carries a feature vector $x \in \mathbb{R}^3$, and Pearson $|r|$ edge weights enter attention as edge attributes. Layer 1 is GATConv(3, 16, heads = 4, edge_dim = 1), giving a 64-dimensional output followed by LayerNorm, ELU and dropout (p = 0.2); Layer 2 is GATConv(64, 16, heads = 4, edge_dim = 1) followed by LayerNorm. A linear projection of the raw features is added to the Layer-2 output (residual connection) before a final ELU and dropout, giving the 64-dimensional gene embedding h_gene used for both risk prediction and pathway scoring. Mean-pooling h_gene over nodes gives a patient embedding $z \in \mathbb{R}^{64}$, which is concatenated with the three clinical covariates and passed through

a two-layer perceptron (67 → 32 → 1, ELU, dropout 0.2) to yield a scalar Cox risk score. The model has 7,585 trainable parameters; patients sharing a fold's topology are processed as one block-diagonal batched graph, which is numerically equivalent to processing each graph independently.

Training and evaluation

Cox-GAT was trained with the Cox partial-likelihood loss and Breslow tie handling[19]:
L = −(1/n_events) * sum over i with event_i=1 of [r_i − log(sum over j with t_j >= t_i of exp(r_j))]
computed with a numerically stable log-cumulative-sum-exp over risk-sorted patients. Optimization used AdamW[20] (learning rate 1×10^{-3}, weight decay 1×10^{-4}) with a 10-epoch linear warmup and cosine decay to 1×10^{-5}, for up to 150 epochs with early stopping on validation C-index (patience 25). Gradients were accumulated to an effective batch size of 16 with norm clipping at 1.0 and mixed precision. Each fold was trained independently on a single NVIDIA RTX 3050 Laptop GPU (4 GB), and the best-validation checkpoint was used for testing.

Four metrics were computed per fold, each implemented from first principles and validated against reference implementations[21,22]: Harrell's concordance index (C-index)[23] (95% bootstrap confidence intervals, 1,000 resamples, for the external ensemble); the integrated Brier score (IBS)[24] with inverse-probability-of-censoring weights; the integrated time-dependent AUC (iAUC) using the cumulative/dynamic estimator[25]; and the log-rank test[26] between high- and low-risk groups, with the threshold fixed at the training fold's median risk score. Baseline hazards and censoring distributions were estimated on the training split only. This fit-on-train-only convention is the central methodological safeguard of the study.

Pathway dysregulation scoring

For each of the 15 fold checkpoints, a forward pass on that fold's test patients retained the 64-dimensional gene embeddings h_gene rather than only the scalar risk. For each of the 805 Reactome pathways with at least three mapped PPI nodes, a pathway dysregulation score was defined as the L2 norm of the mean-pooled gene embedding over the pathway's member nodes (a statistic developed for this study). Because each patient is in the test set once per seed, scores were averaged over the three appearances, giving a 491 × 805 patient-by-pathway matrix.

Pathways were ranked by IQR × Δ, the product of the inter-quartile range of the pathway score across patients and the absolute difference in mean score between risk groups split at the median aggregated risk (a study-specific heuristic). The top 50 pathways were each tested by univariate Cox regression[19] on the standardized score (lifelines[22], penalizer 0.1), with Benjamini–Hochberg correction[15] across the 50 tests. Thirty-one pathways (62%) were significant at $q < 0.05$, exceeding the pre-specified target of at least five; the ten most significant are reported in Table 1. Because ranking used the model's own predicted risk groups and correction covers only 50 of 805 pathways, P values are exploratory. The L2-norm score is unsigned, so hazard ratios describe the association of embedding magnitude with risk, not the direction of pathway regulation.

Pathway name	PPI nodes	HR (95% CI)	P	q
Regulation of cholesterol biosynthesis by SREBP (SREBF)	6	0.68 (0.59–0.78)	6.39×10^{-8}	0.000003
SUMOylation	32	1.39 (1.21–1.59)	2.57×10^{-6}	0.000064
Citric acid cycle (TCA cycle)	6	0.72 (0.62–0.83)	8.58×10^{-6}	0.000143
Regulation of MITF-M-dependent genes involved in cell cycle and proliferation	5	1.34 (1.17–1.53)	2.33×10^{-5}	0.000292
Transcriptional regulation by RUNX3	13	1.31 (1.14–1.51)	1.23×10^{-4}	0.000939
Aerobic respiration and respiratory electron transport	28	0.74 (0.64–0.86)	1.29×10^{-4}	0.000939
Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S	12	1.30 (1.14–1.48)	1.31×10^{-4}	0.000939

Pathway name	PPI nodes	HR (95% CI)	P	q
TFAP2 (AP-2) family regulates transcription of growth factors and their receptors	7	1.27 (1.11–1.44)	4.00×10^{-4}	0.002348
Degradation of beta-catenin by the destruction complex	5	1.28 (1.12–1.47)	4.71×10^{-4}	0.002348
Influenza Infection	28	1.27 (1.11–1.45)	5.83×10^{-4}	0.002348

Table 1 — Top 10 prognostic Reactome pathways ranked by Benjamini–Hochberg-adjusted P value, with ties ordered by raw P (univariate Cox regression of the standardized pathway score, 491 patients).
HR, hazard ratio per 1 SD increase in the pathway dysregulation score; CI, confidence interval; PPI nodes, number of pathway member genes mapped onto the PPI graph; q, Benjamini–Hochberg-adjusted P value across the 50 pathways tested. Cox models used a ridge penalizer of 0.1.

Gradient-based within-pathway biomarker attribution

For each of the 31 significant pathways, gene-level importance was estimated by gradient-based input attribution[27] on a single representative fold (seed42_fold0, chosen for computational tractability; not cross-validated, see Discussion). The squared L2 norm of each pathway's mean-pooled embedding was back-propagated to the input node features for 30 high-risk test patients (risk at or above the aggregated median), and gradient magnitudes were averaged across patients. A pooled ranking took each gene's maximum score across its pathways (gene score), covering 230 unique genes. To assess biological plausibility, the ten top-ranked pathways (Table 1) and the four genes highlighted in the Abstract were compared with published ccRCC and renal cell carcinoma studies through a targeted, non-systematic literature search carried out after pathway and gene selection (results in the Discussion).

External validation on an independent CPTAC-3 cohort

External generalization was tested on the CPTAC-3 ccRCC cohort[8,28], processed independently of TCGA-KIRC. After restricting to "Renal cell carcinoma, NOS" histology (236 patients) and one tumor sample per patient (223 patients), 188 patients with complete RNA, CNV and survival data were retained. To prevent leakage, the CPTAC gene panel was fixed to the TCGA-derived PPI gene index (no differential-expression discovery was performed on CPTAC), and CPTAC RNA counts were variance-stabilized blind (PyDESeq2, use_design=False)[13] with the same settings as TCGA and without reference to any outcome. CPTAC CNV was already GISTIC2-discrete on the same −2 to 2 scale.

Each of the 15 TCGA-trained models was applied to CPTAC without refitting, using its own fold's frozen training-tumor statistics, PPI edge set, baseline hazard, censoring distribution and risk-group threshold. Genes in the TCGA panel but absent from CPTAC were imputed to the training-fold mean (RNA) or 0 (CNV). An ensemble risk score per patient was obtained by averaging the 15 fold-level scores, and patients were split at the median ensemble score into equal high- and low-risk groups (94 each).

Baseline architectures and ablation design

Three reference architectures were re-implemented and trained from scratch on the same 15 folds with identical features, Cox objective, optimizer schedule and early-stopping rule, matched to Cox-GAT in depth: (i) DeepSurv[29], a multilayer perceptron on the same per-gene features without message passing, which isolates graph structure from depth; (ii) GCN-surv, a two-layer graph convolutional network[30] on the identical PPI topology and edge weights, which isolates attention; and (iii) GraphSAGE-surv, a two-layer GraphSAGE network[31] on the same graph, included because GraphSAGE was the best graph architecture on TCGA-KIRC in an independent benchmark[32]. Values for other published methods were taken from the original publications and not reproduced.

Nine single-component ablations were trained on the same 15 folds, each differing from the full model in one respect: feature channels (removing CNV, node degree or clinical covariates, or RNA-only input); graph structure (removing all PPI edges, or randomly rewiring the topology at identical edge count and weight distribution); architecture (removing the residual connection or using one attention layer); and training recipe (removing dropout). A tenth variant, sum-pooling instead of mean-pooling, was a readout-scale check rather than a

structural ablation, since with a fixed node count it only rescales the pooled embedding. Effects are reported as changes in mean test C-index, IBS and iAUC relative to a re-trained full model.

Software and data availability

Analyses used Python 3.12 with PyDESeq2, PyTorch Geometric, lifelines[22] and scikit-survival[21]. Every pipeline stage wrote a machine-readable audit record to a versioned log, and automated leakage checks had to pass before downstream stages ran. Seeds were fixed throughout (master seed 42; CV seeds 42, 123, 456).

Data and code availability. The TCGA-KIRC RNA-seq, copy-number and clinical data analyzed in this study are publicly available from the Genomic Data Commons (https://portal.gdc.cancer.gov) and cBioPortal (https://www.cbioportal.org). The latest CPTAC-3 clear cell renal cell carcinoma data release, as analyzed in this study, is publicly available through the NCI Genomic Data Commons and the associated publications[9,33]. The source code, trained fold models, fold assignments and audit logs supporting the findings of this study are available at [https://github.com/jiyad99/cancer-biomarker-gnn.git].

Results

Cox-GAT achieved good discrimination with moderate fold-to-fold variability

Across the 15 cross-validation folds, no patient overlapped between partitions and all automated audits passed. Cox-GAT achieved a mean test C-index of 0.7697 ± 0.0458 (range 0.6323–0.8170), an integrated Brier score of 0.1615 ± 0.0332 and an iAUC of 0.7913 ± 0.0494. Log-rank stratification was significant ($P < 0.05$) in all 15 folds, including the weakest (seed42_fold3, C-index 0.632; $P = 0.0098$). Aggregated out-of-fold risk groups, split at the median (246 high-risk and 245 low-risk patients), differed significantly (log-rank $P = 1.1 \times 10^{-18}$).

Pathway scoring identified 31 prognostic Reactome pathways

Of the 807 non-redundant Reactome pathways, 805 retained at least three PPI nodes and were scored across all 491 patients. Univariate Cox regression of the 50 top-ranked pathways identified 31 (62%) significant at Benjamini–Hochberg $q < 0.05$ (Table 1 lists the ten most significant). Metabolic pathways (cholesterol biosynthesis/SREBP, TCA cycle, and aerobic respiration and respiratory electron transport) carried protective hazard ratios (HR 0.68–0.74), whereas transcriptional, translational and proliferative pathways (SUMOylation, MITF-M-dependent genes, RUNX3, TFAP2, the β-catenin destruction complex and cap-dependent translation initiation), together with influenza infection, were associated with increased risk (HR 1.27–1.39).

Gradient attribution nominated candidate prognostic genes

Gradient attribution across the 31 significant pathways, computed on one representative fold with 30 high-risk test patients, spanned 230 unique genes. The highest pooled gene scores were obtained for ANGPTL4 (28.97), CYP51A1 (17.41), CGA (16.24), KCTD1 (13.06) and PTK6 (11.56); AXIN2 and ALDOB were also among the ten highest-ranked genes.

Cox-GAT generalized to CPTAC-3 without refitting

Applying the 15 TCGA-trained fold models to CPTAC-3 (n = 188), with all statistics frozen from TCGA training folds, gave a mean per-fold external C-index of 0.7032 ± 0.0354 (range 0.6403–0.7670), IBS of 0.0852 ± 0.0047 and iAUC of 0.7304 ± 0.0342, with significant log-rank separation in all 15 folds. The lower external than internal IBS (0.0852 versus 0.1615) should not be read as better calibration, because the IBS depends on event rate and follow-up horizon, which differ between cohorts. The ensemble risk score achieved a C-index of 0.707 (95% CI 0.598–0.805) and divided 94 high-risk and 94 low-risk patients into groups with significantly different survival (log-rank $P = 0.00809$).

Cox-GAT shows a modest advantage over baselines

Under an identical protocol, Cox-GAT achieved the highest mean test C-index (0.7697 ± 0.0458), ahead of GCN-surv (0.7610 ± 0.0518), GraphSAGE-surv (0.7606 ± 0.0549) and DeepSurv (0.7578 ± 0.0560), together with the lowest IBS (0.1615) and the highest iAUC (0.7913) of the four models. It reached significant log-rank stratification in all 15 folds, compared with 14 of 15 for DeepSurv. The margins over DeepSurv (ΔC-index = +0.0119; graph-free model of matched depth) and over GCN-surv (+0.0087; the non-attention graph model) are small relative to the fold-to-fold standard deviation and no paired significance test was performed, so they indicate at most a modest advantage in mean performance rather than a demonstrated effect of either the graph or attention.

For context only, the original authors report discrete-time C-index values on TCGA-KIRC (n = 313, patient-similarity graph) of 0.7099 for GraphSAGE-surv, 0.7077 for GCN-surv, 0.6962 for GAT-surv and 0.6581 for MLP-surv[32], and 0.745 for the multi-omics graph convolutional network FGCNSurv[33]. These come from different cohorts and protocols and are not a controlled benchmark. Because Cox-GAT also uses clinical covariates, and

its no-clinical ablation (C-index 0.6575) falls below every one of them, its higher internal C-index (0.7697) should not be read as evidence that it outperforms these methods.

Clinical covariates and the residual connection contribute most; random rewiring hurts

Ten variants (nine ablations and the sum-pooling check) were trained on the identical 15 folds. The re-trained full model reproduced the original result (C-index 0.7698 versus 0.7697), so differences reflect the ablated component rather than retraining noise. Excluding the sum-pooling check, the five largest reductions in mean test C-index came from removing the clinical covariates (ΔC-index = −0.1123, to 0.6575), the residual connection (−0.0233), randomly rewiring the PPI graph (−0.0205), removing node degree (−0.0154) and removing the CNV channel (−0.0115). The roles of node degree and CNV are not established, because the RNA-only variant (0.7619), which removes both, outperformed either single removal (0.7544 and 0.7583). Removing the clinical covariates produced by far the largest loss, indicating that age, stage and grade carry prognostic information not recovered from the expression graph. Randomly rewiring the PPI graph hurt, whereas removing all PPI edges cost only 0.0065 (0.7633 versus 0.7698); the real STRING graph therefore beat no graph by an untested margin that is small relative to the fold-to-fold standard deviation, so a benefit of the true topology is not established, although a wrongly wired graph is worse than none.

Sum-pooling produced the largest degradation of all variants (ΔC-index = −0.1641; IBS +0.1118). Because the graph has a fixed number of nodes (4,414), it merely rescales the pooled embedding and shifts the input scale of the risk head, so this most plausibly reflects readout-scale sensitivity rather than information loss; it is not counted among the component contributions and supports mean-pooling as the readout.

Discussion

Enforcing fold-isolated preprocessing at every stage still yielded good internal discrimination (C-index 0.7697) with moderate fold-to-fold variability. Because no non-isolated pipeline was run for comparison, this study cannot quantify how much leakage would have inflated performance; it shows only that strict isolation is compatible with good discrimination.

External validation provides the strongest evidence of generalizability. The decline from the internal C-index (0.7697) to the CPTAC-3 ensemble C-index (0.707) is consistent with differences between cohorts in patient population, sample handling and copy-number calling, and no fold had access to any CPTAC statistic during training. That the ensemble retained significant stratification (log-rank P = 0.00809) without refitting addresses a gap noted in the Introduction; the wide confidence interval (0.598–0.805) reflects the modest cohort size (n = 188).

The comparisons and ablations qualify how this performance should be interpreted. The advantage over graph-free DeepSurv (+0.0119) and GCN-surv (+0.0087) was small relative to the fold-to-fold standard deviation (0.046–0.056), so the graph architecture, and attention in particular, provides at most a modest gain. By contrast, removing the clinical covariates lowered the C-index by 0.112, roughly five times more than any other structural ablation, indicating that most discriminative performance derives from age, stage and grade rather than from the molecular graph. Whether the true PPI topology adds information beyond the node features is not established, since the real STRING graph exceeded no graph by only 0.0065 (untested). The pathway- and gene-level findings should be read in light of this dominant clinical contribution.

These findings align with established ccRCC biology. VHL loss activates HIF signaling and drives a Warburg-like metabolic shift that includes down-regulation of TCA-cycle genes[2]. Consistent with this, the TCA-cycle and aerobic-respiration pathway scores carried protective hazard ratios (0.72 and 0.74), and the metabolic genes CYP51A1 and ALDOB were among the top-ranked attributed genes. The cholesterol-biosynthesis/SREBP pathway showed the strongest single association (HR 0.68), driven by CYP51A1; because the prognostic direction of cholesterol-pathway activity in ccRCC is less firmly established than that of the TCA/HIF axis, this finding is a candidate for follow-up. ALDOB down-regulation is linked to advanced stage and worse survival[34], and ANGPTL4, a HIF-inducible gene up-regulated in ccRCC[35], ranked highest overall, although its prognostic value is not established. Because the pathway score is an unsigned norm, these associations describe embedding magnitude rather than direction of regulation, and their alignment with VHL–HIF biology is interpretive.

To place the pathway- and gene-level findings in context, we compared them with published studies in ccRCC and renal cell carcinoma (targeted, non-systematic search; see Methods). Of the ten top-ranked pathways in Table 1, six have previously been linked to renal cancer prognosis or progression: the TCA cycle and aerobic respiration, through the metabolic reprogramming described in the TCGA ccRCC landscape[2]; SUMOylation, whose regulators form prognostic signatures in TCGA-KIRC[36] and underlie a three-gene SUMOylation risk model validated across five cohorts including CPTAC[37]; RUNX3, whose reduced expression is associated with shorter survival and metastasis in renal cell carcinoma[38]; TFAP2/AP-2, for which AP-2α deregulation is associated with

renal cell carcinoma survival[39]; and SREBP-dependent cholesterol and lipid metabolism, which is HIF-2α-driven and associated with poorer survival in ccRCC[40]. A seventh pathway, the MITF-dependent pathway, has experimental support for promoting ccRCC growth and invasion[41], although we found no survival association. The remaining three pathways (translation initiation, the β-catenin destruction complex and Influenza Infection) had no directly relevant ccRCC prognostic report in our search and should be regarded as novel candidates. At the gene level, three of the four genes highlighted in the Abstract have prior support: ALDOB down-regulation is associated with advanced stage and poorer survival[34], ANGPTL4 is up-regulated in ccRCC[35], and CYP51A1 is down-regulated in tumors and was the top-contributing feature of a published prognostic signature[42]. Of these, only ALDOB and CYP51A1 have prognostic evidence, and no ccRCC-specific report was identified for AXIN2. Two directional discrepancies should be noted. Low RUNX3 expression is unfavorable in the literature[38], whereas the RUNX3 pathway score was a risk factor here (HR 1.31); and high SREBP-pathway activity has been linked to worse survival[40], whereas the cholesterol-biosynthesis/SREBP score was protective (HR 0.68). Because the pathway score is an unsigned embedding magnitude, these comparisons concern association with ccRCC biology and not direction of regulation. Several supporting studies are themselves computational or single-centre, so this concordance is supportive but not confirmatory.

This study has several limitations. First, 56% of DEGs lacked a qualifying STRING partner and were excluded from the graph, so their prognostic signal was unavailable to the model. Second, gradient attribution was computed on one representative fold, was not cross-validated, and attributes the norm of the pathway embedding rather than survival risk itself, so the gene ranking is a candidate list rather than a set of established drivers. The pathway ranking and pooled Kaplan–Meier stratification used the aggregated out-of-fold median risk and were not fitted exclusively on training-fold data; because pathways were pre-selected using the model's own risk groups and corrected for only 50 tests, their P values are exploratory. The literature comparison was targeted rather than systematic, and three pathways and one highlighted gene (AXIN2) had no directly relevant ccRCC report identified. Differential expression (569 raw samples) and VST (565 aligned samples) were performed once on the full cohort; only the tumor-versus-normal label, never survival, was used, but held-out patients contributed to both. Third, the margins over the baselines were not tested for significance; classical survival models (penalized Cox regression, random survival forests) and a clinical-covariate-only Cox model were not benchmarked, so the incremental value of the molecular graph over age, stage and grade has not been directly quantified. Finally, TCGA-KIRC and CPTAC-3 are retrospective cohorts, and prospective or multi-institutional validation is required before clinical application.

In conclusion, Cox-GAT, a leakage-audited graph attention network, achieved a mean internal concordance index of 0.7697 across 15 cross-validation folds on TCGA-KIRC and generalized to an independent CPTAC-3 cohort with an ensemble concordance index of 0.707 without refitting. It identified 31 prognostic Reactome pathways, led by cholesterol biosynthesis, SUMOylation and the TCA cycle, and nominated candidate prognostic genes. Prior renal cancer literature supports six of the ten top pathways and three of the four highlighted genes, but these findings remain exploratory. Under identical folds, Cox-GAT showed a small advantage over DeepSurv, GCN-surv and GraphSAGE-surv, and ablations showed that clinical covariates and the residual connection contribute most to performance, whereas the real PPI graph exceeded no graph by only 0.0065 (untested). Rigorous, audited leakage control is thus compatible with externally validated, biologically interpretable multi-omics survival prediction.

Acknowledgements

The authors received no specific funding for this work.

Conflict of interest

The authors declare no conflict of interest.

References

1. Rose TL, Kim WY. Renal Cell Carcinoma: A Review. JAMA. 2024;332(12):1001-1010. DOI: 10.1001/jama.2024.12848
2. The Cancer Genome Atlas Research Network. Comprehensive Molecular Characterization of Clear Cell Renal Cell Carcinoma. Nature. 2013;499(7456):43-49. DOI: 10.1038/nature12222
3. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018;173(2):400-416.e11. DOI: 10.1016/j.cell.2018.02.052

4. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph Attention Networks. In: Proceedings of the 6th International Conference on Learning Representations; 2018 Apr 30-May 3; Vancouver, Canada. OpenReview.net; 2018.
5. Yan H, Weng D, Li D, Gu Y, Ma W, Liu Q. Prior Knowledge-Guided Multilevel Graph Neural Network for Tumor Risk Prediction and Interpretation via Multi-Omics Data Integration. Briefings in Bioinformatics. 2024;25(3):bbae184. DOI: 10.1093/bib/bbae184
6. Liang B, Gong H, Lu L, Xu J. Risk Stratification and Pathway Analysis Based on Graph Neural Network and Interpretable Algorithm. BMC Bioinformatics. 2022;23(1):394. DOI: 10.1186/s12859-022-04950-1
7. Whalen S, Schreiber J, Noble WS, Pollard KS. Navigating the Pitfalls of Applying Machine Learning in Genomics. Nature Reviews Genetics. 2022;23(3):169-181. DOI: 10.1038/s41576-021-00434-9
8. Clark DJ, Dhanasekaran SM, Petralia F, Pan J, Song X, Hu Y, et al. Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. Cell. 2019;179(4):964-983.e31. DOI: 10.1016/j.cell.2019.10.007
9. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. New England Journal of Medicine. 2016;375(12):1109-1112. DOI: 10.1056/NEJMp1607591
10. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery. 2012;2(5):401-404. DOI: 10.1158/2159-8290.CD-12-0095
11. Seal RL, Braschi B, Gray K, Jones TEM, Tweedie S, Haim-Vilmovsky L, et al. Genenames.org: The HGNC Resources in 2023. Nucleic Acids Research. 2023;51(D1):D1003-D1009. DOI: 10.1093/nar/gkac888
12. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G. GISTIC2.0 Facilitates Sensitive and Confident Localization of the Targets of Focal Somatic Copy-Number Alteration in Human Cancers. Genome Biology. 2011;12(4):R41. DOI: 10.1186/gb-2011-12-4-r41
13. Muzellec B, Teleńczuk M, Cabeli V, Andreux M. PyDESeq2: A Python Package for Bulk RNA-seq Differential Expression Analysis. Bioinformatics. 2023;39(9):btad547. DOI: 10.1093/bioinformatics/btad547
14. Love MI, Huber W, Anders S. Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2. Genome Biology. 2014;15(12):550. DOI: 10.1186/s13059-014-0550-8
15. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B. 1995;57(1):289-300. DOI: 10.1111/j.2517-6161.1995.tb02031.x
16. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, et al. The STRING Database in 2023: Protein-Protein Association Networks and Functional Enrichment Analyses for Any Sequenced Genome of Interest. Nucleic Acids Research. 2023;51(D1):D638-D646. DOI: 10.1093/nar/gkac1000
17. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The Reactome Pathway Knowledgebase 2022. Nucleic Acids Research. 2022;50(D1):D687-D692. DOI: 10.1093/nar/gkab1028
18. Fey M, Lenssen JE. Fast Graph Representation Learning with PyTorch Geometric. In: ICLR Workshop on Representation Learning on Graphs and Manifolds; 2019 May 6; New Orleans, USA. 2019.
19. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society Series B. 1972;34(2):187-220. DOI: 10.1111/j.2517-6161.1972.tb00899.x
20. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. In: Proceedings of the 7th International Conference on Learning Representations; 2019 May 6-9; New Orleans, USA. OpenReview.net; 2019.
21. Pölsterl S. scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn. Journal of Machine Learning Research. 2020;21(212):1-6.
22. Davidson-Pilon C. lifelines: Survival Analysis in Python. Journal of Open Source Software. 2019;4(40):1317. DOI: 10.21105/joss.01317
23. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the Yield of Medical Tests. JAMA. 1982;247(18):2543-2546. DOI: 10.1001/jama.1982.03320430047030
24. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and Comparison of Prognostic Classification Schemes for Survival Data. Statistics in Medicine. 1999;18(17-18):2529-2545. DOI: 10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5
25. Uno H, Cai T, Tian L, Wei LJ. Evaluating Prediction Rules for t-Year Survivors with Censored Regression Models. Journal of the American Statistical Association. 2007;102(478):527-537. DOI: 10.1198/016214507000000149
26. Mantel N. Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration. Cancer Chemotherapy Reports. 1966;50(3):163-170.
27. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In: Workshop at the 2nd International Conference on Learning Representations; 2014 Apr 14-16; Banff, Canada. 2014.

28. Li Y, Lih TM, Dhanasekaran SM, Mannan R, Chen L, Cieslik M, et al. Histopathologic and Proteogenomic Heterogeneity Reveals Features of Clear Cell Renal Cell Carcinoma Aggressiveness. Cancer Cell. 2023;41(1):139-163.e17. DOI: 10.1016/j.ccell.2022.12.001
29. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: Personalized Treatment Recommender System Using a Cox Proportional Hazards Deep Neural Network. BMC Medical Research Methodology. 2018;18(1):24. DOI: 10.1186/s12874-018-0482-1
30. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. In: Proceedings of the 5th International Conference on Learning Representations; 2017 Apr 24-26; Toulon, France. OpenReview.net; 2017.
31. Hamilton WL, Ying R, Leskovec J. Inductive Representation Learning on Large Graphs. In: Advances in Neural Information Processing Systems 30; 2017 Dec 4-9; Long Beach, USA. Red Hook: Curran Associates; 2017. p. 1024-1034.
32. Kim SY. GNN-Surv: Discrete-Time Survival Prediction Using Graph Neural Networks. Bioengineering. 2023;10(9):1046. DOI: 10.3390/bioengineering10091046
33. Wen G, Li L. FGCNSurv: Dually Fused Graph Convolutional Network for Multi-Omics Survival Prediction. Bioinformatics. 2023;39(8):btad472. DOI: 10.1093/bioinformatics/btad472
34. Shao Y, Wu B, Yang Z, Liu Z, Ma Y, Huang H, et al. ALDOB Represents a Potential Prognostic Biomarker for Patients with Clear Cell Renal Cell Carcinoma. Translational Andrology and Urology. 2023;12(4):549-571. DOI: 10.21037/tau-22-743
35. Verine J, Lehmann-Che J, Soliman H, Feugeas JP, Vidal JS, Mongiat-Artus P, et al. Determination of Angptl4 mRNA as a Diagnostic Marker of Primary and Metastatic Clear Cell Renal-Cell Carcinoma. PLoS One. 2010;5(4):e10421. DOI: 10.1371/journal.pone.0010421
36. Wu G, Xu Y, Ruan N, Li J, Lv Q, Zhang Q, et al. Genetic Alteration and Clinical Significance of SUMOylation Regulators in Multiple Cancer Types. Journal of Cancer. 2020;11(23):6823-6833. DOI: 10.7150/jca.49042
37. Li SC, Yan LJ, Wei XL, Jia ZK, Yang JJ, Ning XH. A Novel Risk Model of Three SUMOylation Genes Based on RNA Expression for Potential Prognosis and Treatment Sensitivity Prediction in Kidney Cancer. Frontiers in Pharmacology. 2023;14:1038457. DOI: 10.3389/fphar.2023.1038457
38. Chen F, Liu X, Cheng Q, Zhu S, Bai J, Zheng J. RUNX3 Regulates Renal Cell Carcinoma Metastasis via Targeting miR-6780a-5p/E-cadherin/EMT Signaling Axis. Oncotarget. 2017;8(60):101042-101056. DOI: 10.18632/oncotarget.13205
39. Lin PH, Hsieh CH, Yu KJ, Shao IH, Chuang CK, Hsu T, et al. AP-2α Gene Deregulation Is Associated with Renal Cell Carcinoma Patient Survival. BMC Cancer. 2024;24:966. DOI: 10.1186/s12885-024-12526-w
40. Hua X, Ge S, Zhang L, Jiang Q, Chen J, Xiao H, et al. MED15 Is Upregulated by HIF-2α and Promotes Proliferation and Metastasis in Clear Cell Renal Cell Carcinoma via Activation of SREBP-Dependent Fatty Acid Synthesis. Cell Death Discovery. 2024;10:188. DOI: 10.1038/s41420-024-01944-1
41. Kim N, Kim S, Lee MW, Jeon HJ, Ryu H, Kim JM, et al. MITF Promotes Cell Growth, Migration and Invasion in Clear Cell Renal Cell Carcinoma by Activating the RhoA/YAP Signal Pathway. Cancers. 2021;13(12):2920. DOI: 10.3390/cancers13122920
42. Feng T, Wang Y, Zhang W, Cai T, Tian X, Su J, et al. Machine Learning-Based Framework Develops a Tumor Thrombus Coagulation Signature in Multicenter Cohorts for Renal Cancer. International Journal of Biological Sciences. 2024;20(9):3590-3620.DOI: 10.7150/ijbs.94555