G3I-NET: A TRIMODAL ARCHITECTURE FOR TUMOR STAGING AND HOW MUCH ITS GENE GRAPHS ACTUALLY VARY

Anonymous authors
Paper under double-blind review

ABSTRACT

Architectures that condition one modality's structure on another are evaluated by end-task accuracy. That measurement cannot separate a mechanism producing genuinely input-dependent structure from one that has collapsed onto a single solution shared by every input. We build such a mechanism and measure it directly. G3I-Net is a trimodal architecture for non-small-cell lung cancer T-staging: a frozen Swin-UNETR encodes CT volumes, an MLP encodes radiomic descriptors, and a graph neural network encodes pathway genes over a protein–protein interaction prior. A hypernetwork reads the CT embedding and emits one score per gene, from which a per-patient adjacency over that prior is built, so the molecular branch computes over a topology derived from the patient's own tumor rather than one shared across the cohort. On 128 patients with 21 advanced-stage events the model reaches a patient-level AUC of 0.9119, against 0.749 for the strongest classical late-fusion baseline. We then measure the generated graphs. Each sits far from the prior it was built from, so the conditioning is not inert, yet the graphs are nearly identical to one another: the architecture allows sixteen independent ways for patients to differ and the trained model uses about one. Widening the hypernetwork's input leaves that unchanged, and letting it score every edge freely lowers it further. Removing the conditioning entirely leaves every patient's predicted class unchanged, while two implementations of the same model differing only in the order of floating-point operations change two. What limits conditioned graph generation here is not how much the architecture can express, but how much of it the trained model uses.

1 INTRODUCTION

Architectures that make one modality's structure depend on another are evaluated the same way as any other model: by performance on the end task. That measurement cannot separate a conditioning mechanism doing what it was designed to do from one that has collapsed onto a single input-independent solution, because both can reach the same accuracy. Whether such a mechanism works is a different question from whether the model is accurate, and it is rarely asked.

We build one and ask it. G3I-Net is a trimodal architecture for non-small cell lung cancer (NSCLC) T-staging in which a hypernetwork reads a CT embedding and generates a patient-specific gene–gene graph. We then measure what that graph is able to express, what it actually expresses across patients, and how large a difference between architectural variants this pipeline can resolve at all.

T-staging is a useful setting for the question. Stage at diagnosis is the dominant predictor of outcome in NSCLC (Chheang & Brown, 2013; Nam et al., 2024), and two complementary data sources are available: radiomics quantifies tumor shape, intensity and texture from CT (Magnin et al., 2025; Wang et al., 2022), while expression profiling describes the molecular programs driving progression (Fick et al., 2023; Bi et al., 2025). Multimodal models combining them have outperformed single-modality models on several lung cancer tasks (Chen et al., 2025; Vanguri et al., 2022; Wang et al., 2025), but that work targets prognosis and treatment response rather than early- versus advanced-stage classification, and the cohorts are small and single-institution.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The specific gap we address is structural. Gene interaction priors such as protein–protein interaction (PPI) networks enter these models as one fixed graph shared by every patient, which assumes the relevant interactions do not depend on the tumor. Making that graph a function of the imaging phenotype is a natural idea, and it is the idea this paper builds and then audits.

Contributions. We build the architecture, report what it achieves, and then measure whether its central mechanism does what it claims.

1. **A trimodal architecture in which imaging generates molecular topology.** G3I-Net encodes CT volumes with a frozen Swin-UNETR, ten radiomic descriptors with a multilayer perceptron, and 89 pathway genes with a graph neural network over a protein–protein interaction prior. Rather than encoding the three independently and fusing the results, a hypernetwork maps a 256-dimensional CT embedding to a per-patient modulation of that prior, so each patient's expression profile is convolved over a graph built from their own tumor rather than one shared across the cohort. On 128 patients with 21 advanced-stage events the model reaches a patient-level AUC of 0.8972, which a per-patient mixture-of-gates module raises to 0.9119 (95% CI $[0.840, 0.967]$) at 0.571 sensitivity and 0.944 specificity, against 0.749 for the strongest classical late-fusion baseline (§3, §4).
2. **The conditioning is ahead by 0.0085 AUC, and we measure what it actually learns.** Against a static control in which every patient receives the same graph, the patient-specific graph scores $+0.0085$ (95% CI -0.028 to 0.007) and reclassifies none of the 128 patients. Measured directly, each patient's adjacency sits 73% away from the prior, so the mechanism is active rather than inert, yet of the sixteen dimensions of patient-specificity available the realized across-patient family has an effective rank of 1.72. Widening the conditioning input leaves this unchanged, and freeing the mask to score every edge independently lowers it further (§4).
3. **Marginal gene signal is not incremental gene signal.** Every modality subset containing radiomics reaches 0.869 to 0.897 and none differs from the full model; every subset without it falls to 0.661 to 0.677. Gene expression separates stage above chance on its own, yet the same screen that recovers 535 histology-associated genes recovers none for T-stage conditional on imaging (§4).
4. **Nondeterminism exceeds the effects being compared.** Two implementations of the same model, differing only in floating-point accumulation order, reclassify two of 128 patients. Nine gating mechanisms then fail to separate from the ungated model across 52 configurations on the endpoint fixed in advance, and their ordering by AUC reverses under it (§4, §5).

2 Related Work

Multimodal radiogenomics in lung cancer. Models combining imaging with molecular data have been applied across several NSCLC endpoints. Ellen et al. (2023) predict survival from TCGA using an autoencoder over expression and clinical variables, reporting concordance indices of 0.69 and 0.62 for the two major subtypes. Kim et al. (2022) predict post-resection recurrence by ensembling clinical variables with handcrafted and learned radiomic features, outperforming TNM staging alone. Sangeetha et al. (2024) fuse imaging, expression and clinical variables through concatenation with attention-based weighting. These works target prognosis, recurrence and detection; early versus advanced T-stage classification receives little attention, and the cohorts are small and single-institution. Reviews reach the same conclusion (Thawani et al., 2018; Chen et al., 2017), adding that imaging protocols and feature extraction pipelines are not standardised across studies.

Graph priors over molecular data. Where expression is modelled as a graph, the topology is supplied by a curated interaction network and held fixed. Rhee et al. (2018) convolve breast cancer expression over a STRING (Szklarczyk et al., 2023) protein–protein interaction graph for subtype classification, and Schulte-Sasse et al. (2021) integrate multiomics data over PPI networks to identify cancer genes. In both cases the graph is a property of the database rather than of the patient, and every sample is convolved over the same edges. This asserts that which interactions matter does not depend on the tumour, an assumption that is rarely stated and, to our knowledge, not tested.

Conditioned and generated structure. Making one component's parameters a function of the input is standard. Hypernetworks (Ha et al., 2017) generate weights for a target network from a conditioning vector, feature-wise modulation (Perez et al., 2018) conditions activations on an auxiliary input, and gating mechanisms (Hu et al., 2018; Fedus et al., 2022) re-weight representations per example. For graphs, attention re-weights edges within a fixed topology (Veličković et al., 2018), while structure learning replaces the given graph with one optimised for the task (Franceschi et al., 2019; Chen et al., 2020). Those methods infer a single graph for a dataset from the same features the model classifies; here the graph is generated per patient from a different modality. In every case the mechanism is specified, trained end to end, and evaluated by downstream performance, and how much the generated structure actually varies with the input is not reported.

3 Method

3.1 Problem Formulation and Notation

The public NSCLC Radiogenomics cohort (Bakr et al., 2017; 2018; Clark et al., 2013; Gevaert et al., 2012) provides a CT volume with a tumour segmentation and a gene expression profile per patient, from which we derive three inputs: a tumour-centred volume $\mathbf{V} \in \mathbb{R}^{1\times128\times128\times64}$, ten radiomic descriptors $\mathbf{r} \in \mathbb{R}^{10}$ selected by LASSO (§3.3.2), and expression values for 89 pathway genes $\mathbf{g} \in \mathbb{R}^{89}$, the intersection of a curated pathway list with the measured panel. Those genes are the nodes of a fixed protein–protein interaction graph $\mathbf{A}_{\text{static}}$ from STRING (Szklarczyk et al., 2023) with 440 undirected edges.

We predict the T component of the TNM stage, binarised into *Early* (T1/T2) and *Advanced* (T3/T4), over 128 patients of whom 21 are Advanced. Throughout, $d = 64$ is the width every encoder produces, $B = 4$ the batch size, $\sigma(\cdot)$ the logistic sigmoid and $\odot$ the elementwise product.

3.2 Architecture Overview

Three encoders process the CT volume, the radiomic descriptors and the gene expression vector, each producing a feature of width d (Figure 1).

The CT encoder produces two outputs. The first is the d-dimensional feature that enters the fusion; the second is a 256-dimensional embedding that does not, and is instead passed to a hypernetwork that generates a patient-specific adjacency over the gene nodes.

The three features are concatenated into a $3d$-dimensional vector, re-weighted per patient by a mixture-of-gates module, and reduced to the fused width by a gated fusion block. Two heads read the fused representation: a binary T-stage classifier and an auxiliary survival risk head.

The reported model is a single configuration: fused width 64, all three modalities active, patient-specific graph, mixture-of-gates fusion and survival auxiliary. Ablations vary one element at a time.

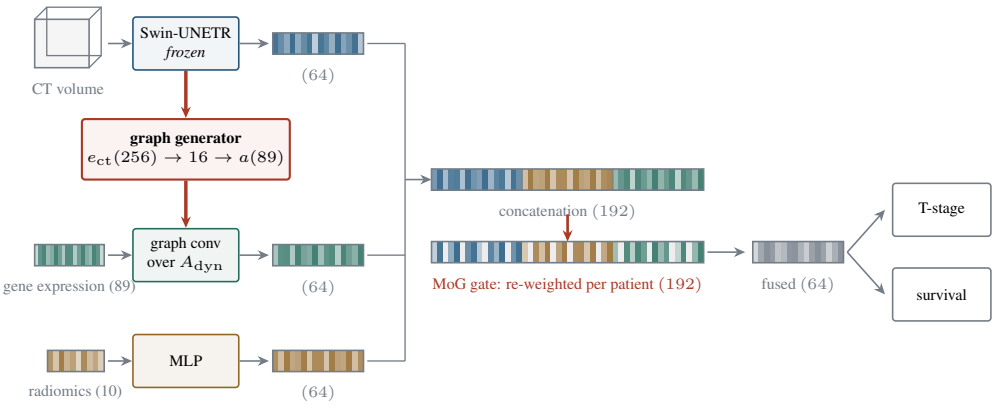

Figure 1: The G3I-Net architecture. Bar length is proportional to tensor dimension. The red path is the contribution: the CT encoder's second projection builds this patient's gene graph.

3.3 MODALITY-SPECIFIC ENCODERS

3.3.1 CT ENCODER

Each patient contributes one tumour-centred volume $\mathbf{V} \in \mathbb{R}^{B\times1\times128\times128\times64}$, single-channel, resampled to 1 mm isotropic spacing and clipped to the $[-1000, 400]$ HU window. The imaging branch is the Swin transformer trunk of a Swin-UNETR (Hatamizadeh et al., 2022) at `feature_size` $= 48$, which reduces the volume to $4\times4\times2$ while growing the channel count from 48 to 768.

We take the trunk's final hidden state $\mathbf{Z} \in \mathbb{R}^{B\times768\times4\times4\times2}$ as the volume representation; the width 768 is fixed by `feature_size` and does not scale with the model's fusion width d. The trunk is initialised from a checkpoint self-supervised on 5,050 unlabelled volumes (Tang et al., 2022) and held frozen, contributing no gradient. Freezing is a constraint rather than a convenience: at 62.2M parameters against 21 positive events, a trainable trunk would exceed the rest of the model by two orders of magnitude, and every comparison between architectural variants would become a comparison between imaging-encoder fits.

The bottleneck is global-average-pooled over its three spatial axes to $\mathbf{z} \in \mathbb{R}^{768}$, retaining channel response and discarding spatial arrangement, and read by two independent $\mathrm{Linear}\to\mathrm{LayerNorm}\to\mathrm{GELU}$ heads—the only trainable parameters in this branch. The first yields $\mathbf{h}_{\mathrm{ct}} \in \mathbb{R}^d$, concatenated with the radiomic and molecular features and therefore constrained to their shared width. The second yields $\mathbf{e}_{\mathrm{ct}} \in \mathbb{R}^{256}$, which does not enter the fusion: it is passed to the hypernetwork of §3.4 and must specify an adjacency over 89 nodes rather than contribute one feature among three, and is wider for that reason.

3.3.2 RADIOMICS ENCODER

We extract radiomic descriptors with PyRadiomics (van Griethuysen et al., 2017) from the same tumour-centred volumes, over the original image and its Laplacian-of-Gaussian and eight-band wavelet transforms, with all texture matrices discretised to 64 bins. This gives 1,037 descriptors per patient, far more than 128 samples can support, so we reduce them in two steps.

First, LASSO regression (Tibshirani, 1996) at $\alpha = 0.005$ ranks the descriptors and drives most coefficients to zero. Conventional shape parameters such as sphericity and maximum 3D diameter are among those eliminated; what survives is dominated by wavelet-filtered first-order and texture statistics. Second, we choose how many to keep by stratified five-fold cross-validation over the top $k \in \{10, 20, 30, 50, 100\}$. Performance peaks at $k = 10$ (AUC 0.713) and falls steadily as more are added (0.680, 0.617, 0.562, 0.505), so descriptors beyond the first ten contribute noise rather than signal at this sample size. The ten we keep are listed in Table 1, and each is standardised by z-score.

The ten values then pass through a two-layer perceptron, $\mathrm{Linear}(10 \to 64) \to \mathrm{GELU} \to \mathrm{Linear}(64 \to d)$, with bias terms and no normalisation or dropout, giving $\mathbf{h}_{\mathrm{rad}} \in \mathbb{R}^d$. At $d = 64$ this is 4,864 parameters, the smallest of the three branches.

	Filter	Descriptor	Coef.
1	Wavelet LHL	`firstorder_Kurtosis`	+0.221
2	Wavelet HLL	`glszm_SmallAreaHighGrayLevelEmphasis`	−0.131
3	LoG σ=3	`glrlm_GrayLevelNonUniformityNormalized`	+0.121
4	Wavelet LLL	`firstorder_RobustMeanAbsoluteDeviation`	+0.115
5	Original	`glszm_SmallAreaEmphasis`	−0.108
6	Wavelet HHH	`firstorder_RootMeanSquared`	+0.107
7	Wavelet HLL	`glszm_GrayLevelNonUniformityNormalized`	−0.096
8	LoG σ=1	`glcm_MaximumProbability`	+0.094
9	Wavelet LLH	`gldm_SmallDependenceHighGrayLevelEmphasis`	+0.090
10	Wavelet HHH	`glszm_SmallAreaEmphasis`	+0.090

Table 1: The ten radiomic descriptors retained by LASSO at $\alpha = 0.005$, ordered by coefficient magnitude. LHL, HLL, LLL, LLH and HHH denote wavelet sub-bands; LoG σ is the Laplacian-of-Gaussian kernel width in millimetres.

3.4 Patient-Specific Gene Graph Generation

The gene branch computes over a graph that depends on the patient. A hypernetwork (Ha et al., 2017) reads the CT embedding and produces the parameters of that graph, and a fixed construction builds it from them.

Emitting the adjacency directly would require $89^2 = 7{,}921$ outputs per patient from a cohort of 128, so the hypernetwork instead emits one score per gene. Three linear layers of widths $256 \to 128 \to 64 \to 16$, with GELU activations and dropout 0.1 after the first, give a modulation code $\mathbf{m} \in \mathbb{R}^{16}$, which expands to $\mathbf{a} = \sigma(\mathbf{W}_{\text{mod}}\mathbf{m}) \in (0,1)^{89}$:

$$\mathbf{M}_{ij} = \sqrt{\mathbf{a}_i \mathbf{a}_j + \epsilon}, \qquad \mathbf{A}_{\text{dyn}} = \frac{\mathbf{A}_{\text{static}} \odot \mathbf{M}}{\sum_k (\mathbf{A}_{\text{static}} \odot \mathbf{M})_{ik} + \epsilon}, \qquad \epsilon = 10^{-8}. \tag{1}$$

Each edge is scaled by the geometric mean of its two endpoint scores. Two properties follow from the construction and hold for every patient. First, the prior is multiplied rather than added to, so an edge absent from STRING can never appear. Second, $\sqrt{\mathbf{a}_i \mathbf{a}_j}$ factors as $\sqrt{\mathbf{a}_i}\sqrt{\mathbf{a}_j}$ and the row normalisation divides by a sum over j, so $\sqrt{\mathbf{a}_i}$ cancels exactly: the graph depends on $\mathbf{a}$ only through a per-column reweighting, and is unchanged if $\mathbf{a}$ is scaled by any positive constant.

Node values are embedded by $\text{Linear}(1 \to d)$ and propagated by two graph convolutional layers (Kipf & Welling, 2017) with residual connections (He et al., 2016):

$$\mathbf{h}^{(l+1)} = \text{ReLU}\Big(\mathbf{A}_{\text{dyn}}\, \mathbf{W}^{(l)} \mathbf{h}^{(l)} + \mathbf{b}^{(l)}\Big) + \mathbf{h}^{(l)}, \qquad l = 0, 1, \tag{2}$$

with dropout 0.3 after each. Nodes are aggregated by attention pooling and the pooled vector is projected by $\text{Linear}(d \to d)$ to $\mathbf{h}_{\text{gene}}$. The width d sets both the fusion width and the internal width of the graph network.

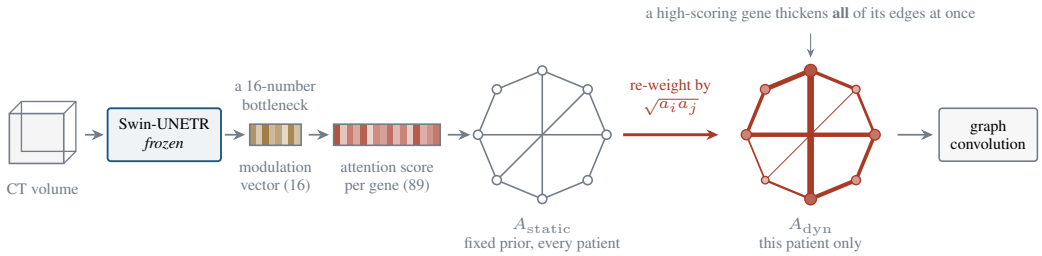

Figure 2: Constructing the patient-specific gene graph. Graphs are schematic with eight nodes rather than 89; node shade and edge thickness show attention.

3.4.1 Capacity and Utilisation of the Conditioning

What this construction permits and what the trained model realises are different quantities. We measure both.

The modulation is rank one for every patient, since $\mathbf{M} = \sqrt{\mathbf{a}}\sqrt{\mathbf{a}}^{\top}$ is an outer product and the ratio of second to first singular value never exceeds 6.4×10^{-6}. A patient can turn all of a gene's edges up or down together, but cannot turn one of them up and another down. The mechanism expresses patient-specific gene importance, not patient-specific gene interaction.

The scores are generated from sixteen numbers, so sixteen directions of variation between patients are available. The realised effective rank is 1.72 (per-fold range 1.09 to 3.01), with 84.9% of the variance on the leading direction. Each patient's graph is therefore close to a single scalar adjustment of one fixed reweighting pattern.

The prior's topology limits propagation independently of the conditioning. Its 440 undirected edges form six disconnected components of sizes $26/23/18/13/8/1$, so with two convolutional layers information never crosses between components, and one gene has no edges at all.

3.5 Per-Patient Mixture-of-Gates Fusion

Which modalities matter may differ between patients, so the fusion is preceded by a gate that is itself a function of the patient. The three encoder outputs are concatenated to $\mathbf{x} \in \mathbb{R}^{3d}$. A router reads that vector and assigns a weight to each of E experts; every expert produces its own multiplicative gate over all $3d$ coordinates, and the gate applied is their weighted average:

$$\mathbf{w} = \mathrm{softmax}(\mathrm{Linear}(3d \to E)(\mathbf{x})/\tau)\,, \tag{3}$$

$$\mathbf{g}_e = \sigma(\mathrm{Linear}(k \to 3d) \circ \mathrm{GELU} \circ \mathrm{Linear}(3d \to k)(\mathbf{x}))\,, \tag{4}$$

$$\mathbf{x}' = \mathbf{x} \odot \textstyle\sum_{e=1}^{E} \mathbf{w}_e \mathbf{g}_e, \tag{5}$$

with $E = 4$, $k = \max(4, 3d/8) = 24$, $\tau = 1$, and every expert active for every patient. Training adds the Switch load-balancing penalty (Fedus et al., 2022) at $\lambda = 0.01$, which discourages the router from using only one expert. At $3d = 192$ the module carries $38{,}500$ parameters.

The experts emit gates rather than feature transformations, so the CT, radiomics and gene blocks remain separable for the fusion that follows. At $E = 1$ the router weight is identically one and Equation 5 becomes $\mathbf{x} \odot \sigma(\mathrm{MLP}(\mathbf{x}))$, which is squeeze-and-excitation (Hu et al., 2018), so that baseline is nested inside this module rather than being a separate architecture.

3.6 Fusion, Prediction Heads and Objective

The gated representation is split back into three d-dimensional blocks and passed to a gated fusion module producing a 64-dimensional joint representation, read by a binary T-stage classifier and a single-output survival risk head. The objective is

$$\mathcal{L} = 0.7\,\mathcal{L}_{\mathrm{CE}}^{\mathrm{smooth}}(y, \hat{y}) + 0.3\,\mathcal{L}_{\mathrm{Cox}}(\hat{h}, t, \delta) + \mathcal{L}_{\mathrm{gate}}, \tag{6}$$

with label smoothing 0.1 on the first term, the Cox partial likelihood in Breslow form on the second, and the load-balancing penalty of Section 3.5 on the third.

3.7 Implementation and Evaluation

The model carries $402{,}177$ trainable parameters, approximately $19{,}150$ per positive event.

Optimisation uses AdamW at learning rate 10^{-3} with weight decay 0.01 under a one-cycle schedule, for up to 100 epochs with early stopping at patience 30 on $0.5 \cdot \mathrm{AUC} + 0.5 \cdot \mathrm{accuracy}$. Batch size is 4, gradients are clipped at norm 1.0, dropout is 0.3, and class imbalance is handled by a weighted random sampler on the T-stage label. Gate parameters occupy a separate optimiser group at learning rate 0.05 with zero weight decay.

Evaluation uses stratified five-fold cross-validation repeated over five seeds, giving 25 runs per configuration. The splits are fixed and reused by every arm, so any two configurations are compared on identical folds. Out-of-fold probabilities are averaged across seeds to give one probability per patient, and confidence intervals come from a bootstrap that resamples patients rather than folds. The primary endpoint was fixed before the experiments were run and is the change in sensitivity at matched specificity; AUC is reported as a descriptive measure.

4 Results

G3I-Net reaches a patient-level AUC of 0.8972 (95% confidence interval 0.8207 to 0.9583) at 0.944 specificity. Adding a mixture-of-experts gate on the pre-fusion concatenation raises this to 0.9119 (0.8398 to 0.9671) at the same specificity, and that gated configuration is the reported model.

This section compares those figures against classical baselines, isolates which modality carries the staging signal, establishes how large a difference the pipeline can resolve at all, and then measures what the gating and conditioning mechanisms actually do.

4.1 Comparison with existing methods

Five classical baselines were trained under the same cross-validation protocol: logistic regression with early and late fusion, XGBoost, random forest, and a multilayer perceptron. G3I-Net outper-

forms all of them, exceeding the strongest by 0.163 AUC (Table 2). The baselines use radiomics and gene expression without the CT volume, although the radiomics features are themselves derived from the CT scan.

	AUC	Accuracy	Specificity
G3I-Net with MoE gate	**0.9119**	**0.883**	0.944
G3I-Net, ungated	0.8972	0.875	0.944
Logistic regression, late fusion	0.7487	0.749	0.827
Logistic regression, early fusion	0.6722	0.722	0.791
XGBoost	0.6411	0.809	0.935
Random forest	0.6406	0.836	0.998
MLP	0.6073	0.764	0.863

Table 2: Comparison with classical baselines. G3I-Net rows are patient-level; baseline rows are fold-level means over 25 runs and exclude the CT volume.

4.2 Modality ablation

All seven modality subsets were trained under the same protocol on the ungated backbone. Every subset containing radiomics is indistinguishable from the full model; every subset without it is significantly worse (Figure 3). Radiomics carries essentially all of the staging signal, as expected for a T category defined by tumour size and local invasion. Gene expression alone reaches 0.667, above chance, yet adds nothing once radiomics is present: no gene survives correction conditional on imaging, while 535 survive under a histology control.

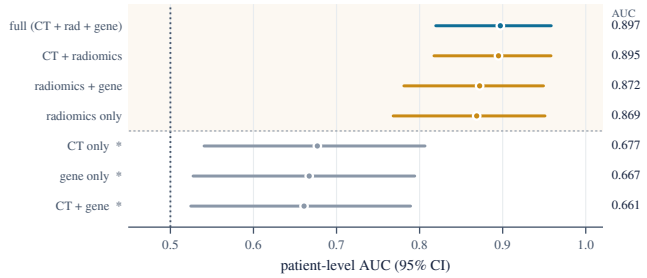

Figure 3: Modality ablation, ungated backbone. Shaded rows contain radiomics; starred arms differ from the full model with an interval excluding zero. Arms without CT have a static gene graph.

4.3 Measurement floor

The reported configuration was trained twice, with the graph modulation computed by two pieces of code implementing the same formula. They are mathematically identical and differ only in the order of the floating-point operations. The results still differ: patient-level AUC changes by 0.0004 (95% confidence interval -0.0124 to 0.0125), two of the 128 patients receive a different prediction, and early stopping selects a different epoch in 12 of the 25 folds. A difference in AUC below about 0.0125 is therefore not distinguishable from running the same model twice, and the comparisons that follow are read against that.

4.4 Gating the fused representation

Nine gating mechanisms were compared at the reported width. The mixture-of-experts gate gives the largest change in AUC, $+0.0147$ (95% confidence interval -0.007 to 0.041), and is the gate carried into the reported model (Figure 4). No arm separates from the ungated backbone on the pre-registered endpoint, and the two orderings disagree: the absolute-value penalty sits below the backbone on AUC yet matches the mixture of gates on sensitivity at matched specificity.

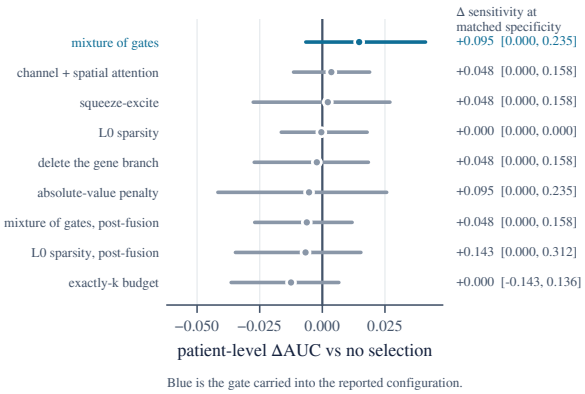

Figure 4: Selection arms at the reported width, patient-level, $n = 128$. Paired change in AUC against the ungated backbone, with the pre-registered endpoint beside it. Blue is the gate carried into the reported model.

4.5 The conditioning mechanism

Section 3.4.1 establishes that the conditioning varies along approximately one direction. Two explanations are available: the rank-one mask restricts what can be expressed, or the hypernetwork receives too little to express it. Both are tested and neither holds. Freeing the mask to score all 440 edges independently lowers the realised rank rather than raising it, and widening the input to include the radiomic features leaves it unchanged ($p = 0.51$). The limit therefore lies upstream of the mask, in the compression of the CT embedding.

Conditioning changes almost no patient-level decisions (Table 3). Static and patient-specific graphs classify all 128 patients identically, and freeing the mask moves one patient, while re-running the same model through a different code path moves two.

	AUC	ΔAUC (95% CI)	Patients reclassified
Reported conditioning (reference)	0.8972	—	—
Same model, different code path	0.8976	$+0.0004$ $(-0.012, 0.013)$	2 / 128
No conditioning (static graph)	0.8887	-0.0085 $(-0.028, 0.007)$	0 / 128
Mask freed to 440 edge scores	0.8741	-0.0231 $(-0.058, 0.002)$	1 / 128

Table 3: Conditioning arms against the ungated backbone. The last column counts patients whose predicted class changes at the reported threshold.

5 Discussion

That the gene branch carries no incremental signal is consistent with T category being an anatomical measurement radiomic descriptors capture directly. The branch is retained because it is the object of study rather than a contributor to accuracy: removing it would remove the mechanism this paper measures.

Why no auxiliary task helps. Gene expression is informative in this cohort, just not about T category: the same screen that recovers 535 histology-associated genes recovers none for T-stage conditional on imaging. A natural response is to supervise the molecular branch on a label it can predict and let the resulting representation serve the primary task. That route is closed here, because no clinical label recorded for this cohort predicts T category.

Nodal status, which clinically tends to track tumour extent, is at 0.508 here. T category is defined by tumour size and local invasion, and in these patients it is statistically isolated from every other

Label	AUC against T-stage	p
Histology (squamous)	0.540	0.46
Overall survival time	0.566	0.34
Recurrence event	0.526	0.63
Overall survival event	0.518	0.76
N-stage	0.508	0.88
M-stage	0.505	0.83

Table 4: Every clinical label available in this cohort, as a predictor of binary T-stage.

recorded endpoint. An auxiliary task can only transfer signal it shares with the primary one, and none is available to share.

Modality importance is a per-patient quantity. A per-patient gate retains all three modalities at near-balanced mass, whereas a single sparsity vector shared across the cohort collapses the model. This is an argument for input-dependent gating independent of whether it improves accuracy.

Limitations. The cohort is single-institution, with 128 patients and 21 advanced cases, and no external validation was performed. The reported configuration was chosen after examining 52 trained configurations on this cohort, so its point estimate is optimistic. Differences below approximately 0.0125 in AUC are not separable from pipeline variance. In some folds the router concentrates almost all weight on one expert, so the gate is then operating as squeeze-and-excitation rather than as a mixture; we therefore do not attribute this arm's result to the mixture mechanism. The survival auxiliary attains a concordance index of 0.505 and contributes no measurable prognostic signal.

Future work. Removing the rank-one constraint without the saturation observed here, by scoring edges under a bounded parameterisation, would separate the constraint from its implementation.

What this establishes. We do not claim that patient-specific gene graphs are unhelpful for staging. We show that this mechanism did not express enough patient-specificity to test the claim, and we report the measurements that establish it. Applying the same architecture to an endpoint where gene expression carries signal, such as histological subtype, would test the conditioning mechanism rather than the availability of molecular information.

6 Conclusion

We presented G3I-Net, a trimodal architecture for NSCLC T-staging in which a hypernetwork generates a patient-specific gene–gene adjacency from the CT embedding, and characterised that mechanism rather than reporting only its accuracy. The architecture permits sixteen independent directions of patient-specific variation; the trained model uses approximately one, and this holds under every parameterisation we tested.

These measurements are taken from trained weights rather than estimated from outcomes, so they do not depend on the size of the cohort. We suggest reporting them alongside accuracy whenever a conditioning mechanism is proposed.

Reproducibility Statement

The dataset is the public NSCLC Radiogenomics collection (Bakr et al., 2017; 2018; Clark et al., 2013). Cross-validation splits are fixed across all experiments: five seeds of stratified five-fold partitions, stored as serialised index files and reused by every arm, so any two arms are compared on identical folds. Preprocessing is specified in Section 3.3: resampling to 1 mm isotropic spacing, clipping to the $[-1000, 400]$ Hounsfield window, and cropping to $128 \times 128 \times 64$ about the tumour. Radiomic descriptors are extracted with PyRadiomics under the settings given there, and the ten retained descriptors are listed in Table 1 with their LASSO coefficients. The gene set and the protein-interaction prior are fixed and are not refitted per fold.

Section 3.7 gives the optimiser, schedule, batch size, early-stopping criterion and parameter counts. Every number reported in this paper is read from a stored result file rather than transcribed, and the figures are generated from those files by scripts that record the checksum of each input they read. Patient-level statistics are computed by pooling out-of-fold probabilities across seeds and resampling patients, using the same procedure for every arm.

Two aspects limit exact reproduction. The Swin-UNETR trunk is initialised from a self-supervised checkpoint that must be obtained separately. Training is not bit-deterministic across code paths: as reported in Section 4.3, two algebraically identical implementations select different early-stopping epochs in 12 of 25 folds, so point estimates are reproducible only to within the measurement floor stated there.

Code and configuration files will be made publicly available.

USE OF LARGE LANGUAGE MODELS

In this work, we used generative AI tools to propose and refine hypotheses, to design and provide feedback on research methodology and experiments, to implement methods, to formulate mathematical claims, and to interpret results. We have not used generative AI tools to generate synthetic data sets, to assist with translation, or to clean and reformat the dataset, and support for qualitative and thematic data analysis is not applicable to this work. Additionally, we used generative AI tools to draft parts of this paper, to edit it for readability, and to create or edit software code and scientific figures.

We have reviewed all AI-assisted work. Every number reported here is read programmatically from a stored result file produced by our own training runs rather than transcribed, and AI-assisted code was verified against the existing implementation before use. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Weiruo Zhang, Ann Leung, Michael Kadoch, Joseph Shrager, Andrew Quon, Daniel Rubin, Sylvia Plevritis, and Sandy Napel. Data for NSCLC radiogenomics collection, 2017. URL `https://www.cancerimagingarchive.net/collection/nsclc-radiogenomics/`. [Data set].

Shaimaa Bakr, Olivier Gevaert, Sebastian Echegaray, Kelsey Ayers, Mu Zhou, Majid Shafiq, Hong Zheng, Jalen Anthony Benson, Weiruo Zhang, Ann NC Leung, et al. A radiogenomic dataset of non-small cell lung cancer. *Scientific Data*, 5:180202, 2018.

Lingfeng Bi, Xin Wang, Jiayi Li, Weimin Li, and Zhoufeng Wang. Epigenetic modifications in early stage lung cancer: pathogenesis, biomarkers, and early diagnosis. *MedComm*, 6(3):e70080, 2025.

Bojiang Chen, Rui Zhang, Yuncui Gan, Lan Yang, and Weimin Li. Development and clinical application of radiomics in lung cancer. *Radiation oncology*, 12(1):154, 2017.

Yu Chen, Lingfei Wu, and Mohammed J Zaki. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 19314–19326, 2020.

Yuanyuan Chen, Dikang Chen, Xiaohui Liu, Hui Jiang, and Xuemei Wang. Deep learning-driven multimodal integration of miRNA and radiomic for lung cancer diagnosis. *Biosensors*, 15(9):610, 2025. doi: 10.3390/bios15090610.

Sophie Chheang and Kathleen Brown. Lung cancer staging: clinical and radiologic perspectives. In *Seminars in interventional radiology*, volume 30, pp. 099–113. Thieme Medical Publishers, 2013.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The

cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013.

Jacob G Ellen, Etai Jacob, Nikos Nikolaou, and Natasha Markuzon. Autoencoder-based multimodal prediction of non-small cell lung cancer survival. *Scientific Reports*, 13(1):15761, 2023.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Cameron N Fick, Elizabeth G Dunne, Manendra B Lankadasari, Brooke Mastrogiacomo, Tetsuhiko Asao, Stijn Vanstraelen, Yuan Liu, Francisco Sanchez-Vega, and David R Jones. Genomic profiling and metastatic risk in early-stage non–small cell lung cancer. *JTCVS open*, 16:9–16, 2023.

Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1972–1982. PMLR, 2019.

Olivier Gevaert, Jiajing Xu, Chuong D. Hoang, Ann N. Leung, Yue Xu, Andrew Quon, Daniel L. Rubin, Sandy Napel, and Sylvia K. Plevritis. Non-small cell lung cancer: identifying prognostic imaging biomarkers by leveraging public gene expression microarray data–methods and preliminary results. *Radiology*, 264(2):387–396, 2012. doi: 10.1148/radiol.12111607.

David Ha, Andrew M. Dai, and Quoc V. Le. HyperNetworks. In *International Conference on Learning Representations (ICLR)*, 2017.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pp. 272–284. Springer, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Gihyeon Kim, Sehwa Moon, and Jang-Hwan Choi. Deep learning with multimodal integration for predicting recurrence in patients with non-small cell lung cancer. *Sensors*, 22(17):6594, 2022.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.

Cheryl Y Magnin, David Lauer, Michael Ammeter, and Janine Gote-Schniering. From images to clinical insights: an educational review on radiomics in lung diseases. *Breathe*, 21(1), 2025.

Brian Nam, Danny Hamm, Nora Katurakes, and Charles Mulligan. Lung cancer screening: Early detection decreases mortality. *Delaware Journal of Public Health*, 10(3):22–24, 2024.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018.

Sungmin Rhee, Seokjun Seo, and Sun Kim. Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 3527–3534, 2018.

SKB Sangeetha, Sandeep Kumar Mathivanan, P Karthikeyan, Hariharan Rajadurai, Basu Dev Shivahare, Saurav Mallik, and Hong Qin. An enhanced multimodal fusion deep learning neural network for lung cancer classification. *Systems and Soft Computing*, 6:200068, 2024.

Roman Schulte-Sasse, Stefan Budach, Denes Hnisz, and Annalisa Marsico. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nature Machine Intelligence*, 3(6):513–526, 2021.

Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L Gable, Tao Fang, Nadezhda T Doncheva, Sampo Pyysalo, et al. The string database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic acids research*, 51(D1):D638–D646, 2023.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20730–20740, 2022.

Rajat Thawani, Michael McLane, Niha Beig, Soumya Ghose, Prateek Prasanna, Vamsidhar Velcheti, and Anant Madabhushi. Radiomics and radiogenomics in lung cancer: a review for the clinician. *Lung cancer*, 115:34–41, 2018.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Joost J. M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina G. H. Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo J. W. L. Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer Research*, 77(21):e104–e107, 2017. doi: 10.1158/0008-5472.CAN-17-0339.

Rami S Vanguri, Jia Luo, Andrew T Aukerman, Jacklynn V Egger, Christopher J Fong, Natally Horvat, Andrew Pagano, Jose de Arimateia Batista Araujo-Filho, Luke Geneslaw, Hira Rizvi, et al. Multimodal integration of radiology, pathology and genomics for prediction of response to pd-(l) 1 blockade in patients with non-small cell lung cancer. *Nature cancer*, 3(10):1151–1164, 2022.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Chunsheng Wang, Congjie Wang, Jianguo Zhang, Mingjun Ding, Yizhi Ge, and Xia He. Development and validation of a radiogenomics prognostic model integrating PET/CT radiomics and glucose metabolism-related gene signatures for non-small cell lung cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, 52(13):4924–4938, 2025. doi: 10.1007/s00259-025-07354-4.

Zixing Wang, Cuihong Yang, Wei Han, Xin Sui, Fuling Zheng, Fang Xue, Xiaoli Xu, Peng Wu, Yali Chen, Wentao Gu, et al. Quantifying lung cancer heterogeneity using novel ct features: a cross-institute study. *Insights into Imaging*, 13(1):82, 2022.