INFORMATION PATHWAYS FOR LEARNING STRUCTURAL PREDICATES ON QUANTUM CIRCUIT DAGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Structural prediction on directed acyclic graphs (DAGs) is an information-access problem before it is an architecture-selection problem. Critical-path membership decomposes exactly into forward longest-path depth d_v, reverse longest-path depth r_v, and a graph-level maximum $L^\star$. Because $d_v + r_v + 1 \leq L^\star$ always holds with equality only at critical nodes, the predicate is a one-sided threshold, which splits the target: within-circuit ranking requires descendant-side path information and provably cannot depend on $L^\star$, while converting that order into a decision requires only $L^\star$. Quantum circuit DAGs make all three quantities deterministically computable, so the split is directly testable. We show that any linear estimate of r_v from quantities a strong baseline already receives is rank-equivalent to forward depth alone, so exact reverse depth supplies a *direction* of information rather than added precision on a quantity already present. Holding a node-wise MLP fixed up to a 385-parameter projection, broadcasting $L^\star$ raises critical-path F1 from 0.573 to 0.826 and adding exact r_v to 0.974, with 0.518, 0.581, 0.846 under a same-generator upper-tail depth shift; macro AUPRC reproduces the ordering (0.832, 0.898, 0.997), so it is not an artifact of a fixed operating point. The identical increment on a message-passing encoder improves the operating point without improving node ranking in distribution while improving both under shift, exactly as the decomposition predicts. Architecture experiments then locate the same quantities in distinct pathways: a DAG-native encoder obtains reverse depth by topological propagation, reaching depth-shift macro AUPRC 0.975 without an explicit broadcast, the evaluated latent-bottleneck Perceiver acquires essentially no rank-relevant descendant information and a $16\times$ latent sweep does not repair this, and a graph transformer is near ceiling with supplied structure but substantially degraded without it. Structural-prediction capability is therefore governed not by architecture labels alone, but by whether the representation obtains reliable access to the information the target demands.

1 INTRODUCTION

Graph representation learning is often framed as architecture selection, choosing among local message passing, latent attention, and full graph attention. For structural prediction on directed acyclic graphs (DAGs), a more fundamental question comes first, *what information must reach a node representation for the target to be computable?*

DAGs make these requirements explicit, since topological depth depends on upstream structure, descendant-derived quantities on downstream structure, and critical-path membership on both together with the graph's longest-path length. Node features, edge orientation, reverse-edge augmentation, message-passing depth, graph-level broadcast, topological propagation, structural attention biases, and global attention therefore constitute distinct information pathways. Quantum circuit DAGs, in which gates become nodes and directed edges encode consecutive use of a shared qubit, make every one of these quantities deterministically computable, and so permit direct verification of which suffice and which pathways provide them, separating information access from label ambiguity. We use critical-path membership as a controlled instrument, not as a claim that learned models should replace dynamic programming given complete DAG access.

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.

Our default inputs deliberately include target-related summaries: exact forward depth, in-degree, out-degree, bucketed descendant count, and sequence position. Feature-enabled performance therefore measures how effectively a model preserves and combines supplied structure rather than recovery from raw connectivity, and reduced-input, edge-direction, receptive-field, and topological-propagation conditions complement it.

The paper makes three contributions, evaluated throughout on a controlled same-generator upper-tail depth shift that stays within observed training path-length support.

A ranking–calibration decomposition of the predicate. Because $d_v + r_v + 1 \leq L^\star$ always holds with equality only at critical nodes, critical-path membership is a one-sided threshold rather than an equality. Ranking nodes within a circuit therefore requires only descendant-side path information and never $L^\star$, while converting that order into a decision requires only $L^\star$. The stronger consequence is that a linear estimate of reverse depth from quantities the baseline already receives is provably rank-equivalent to forward depth alone, so exact reverse depth supplies a direction of information rather than precision on a quantity already available.

Information conditions at fixed architecture. Holding a node-wise architecture fixed up to the 385-parameter projection of Section 3, broadcasting graph-level longest-path length raises standard-set F1 from 0.573 to 0.826, and exact reverse depth raises it to 0.974, with 0.518, 0.581 and 0.846 under depth shift, against all-positive floors of 0.642 and 0.476. The identical increment applied to a message-passing encoder raises standard-set F1 from 0.814 to 0.909, so the effect is not an artifact of the missing graph interaction.

Pathway-level architecture diagnostics. Reverse-edge, parameter-matched depth, latent-count, topological-propagation, and reduced-input experiments expose distinct pathways for descendant communication, local reach, latent routing, and graph-aware global interaction. A DAG-native encoder reaches depth-shift macro AUPRC 0.975 with no explicit broadcast, exceeding a node-wise model handed exact reverse depth and $L^\star$.

Why the diagnosis matters more than the predicate. Compilation and scheduling for quantum hardware are increasingly learned rather than computed, because the quantities that matter, post-transpilation depth, schedule length, and routing overhead, have no closed form. All of them depend on ancestor-side, descendant-side, and graph-level structure, and none permits checking whether a representation actually obtained it. Critical-path membership does. It is solved exactly by a two-pass dynamic program, so its information requirements can be written down and then verified, and failure is attributable to access rather than to label noise or an unidentified dependency. Used this way it shows that failures are not interchangeable. A missing graph-level constant costs 385 parameters to repair and no amount of message passing supplies it for free, whereas a missing descendant-side quantity yields only to propagation along the partial order. And the obvious diagnostic, regressing the suspected-missing quantity from the features already present and substituting the estimate, is a trap that is rank-degenerate here, so $R^2 = 0.845$ coexists with macro-F1 0.026.

2 Circuit DAGs and Information Requirements

2.1 Circuit representation

We represent a quantum circuit as $G = (V, E, X)$, where V is the set of gate nodes, $E \subseteq V \times V$ contains an edge (u, v) when v is the next gate acting on a qubit used by u, and X contains gate type, gate family, and primary-wire information. We write $\mathrm{pa}(v)$ and $\mathrm{ch}(v)$ for the parents and children of v, $\mathrm{desc}(v)$ for the set of descendants of v, and $\bigoplus$ for a permutation-invariant aggregation operator. Figure 1 illustrates the construction.

2.2 Path quantities and critical-path membership

Let path length be counted in nodes, so that a single node is a path of length one. For a node v, let $\ell_{\rightarrow}(v)$ denote the number of nodes on a longest directed path ending at v, and $\ell_{\leftarrow}(v)$ the number on a longest directed path starting at v, each counting v itself. Sources therefore satisfy $\ell_{\rightarrow}(v) = 1$ and sinks $\ell_{\leftarrow}(v) = 1$. We write

$$L^\star = \max_{u \in V} \ell_{\rightarrow}(u) = \max_{u \in V} \ell_{\leftarrow}(u)$$

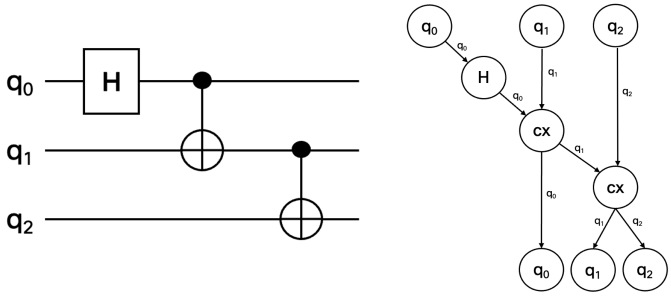

Figure 1: Quantum circuit DAG construction. Gates become nodes, and a directed edge identifies the next operation on a shared qubit. Gates on disjoint qubits may occupy parallel branches, so a node's forward depth d_v is determined by its ancestors, its reverse depth r_v by its descendants, and $L^\star$ by the graph as a whole.

for the number of nodes on a longest directed path in G.

Stored features are zero-indexed, $d_v = \ell_\rightarrow(v) - 1$ and $r_v = \ell_\leftarrow(v) - 1$, so d_v is the forward topological depth of v and r_v its reverse depth, each measured in edges.

Let $\mathrm{CP}(v) \in \{0, 1\}$ indicate whether v lies on at least one longest directed path of G; we call such nodes critical. Because every gate contributes exactly one node, this is the unweighted critical path of the circuit DAG, with no gate-latency weights. Concatenating a longest path ending at v with a longest path starting at v yields a directed path through v on $\ell_\rightarrow(v) + \ell_\leftarrow(v) - 1$ nodes, counting v once, and no longer path through v exists. Hence

$$\mathrm{CP}(v) = 1 \quad \Longleftrightarrow \quad \ell_\rightarrow(v) + \ell_\leftarrow(v) - 1 = L^\star \quad \Longleftrightarrow \quad d_v + r_v = L^\star - 1. \tag{1}$$

Equation 1 is the object the rest of the paper studies: membership is fixed by exactly three quantities, the ancestor-derived d_v, the descendant-derived r_v, and the graph-level $L^\star$. Writing $s_v = d_v + r_v + 1$ for the node count of the longest path through v, the same argument gives

$$s_v \;\le\; L^\star \qquad \text{for every } v \in V, \tag{2}$$

with equality if and only if $\mathrm{CP}(v) = 1$. We verify Equation 2 and its label agreement on all $5{,}985$ standard-set and $367{,}290$ depth-shift nodes: zero violations and zero disagreements. Equation 1 is therefore equivalent to the *one-sided threshold* $d_v + r_v \ge L^\star - 1$, and with $\mathrm{slack}_v = L^\star - s_v \ge 0$ criticality is exactly $\mathrm{slack}_v = 0$.

This separates the target into two competencies. *Ranking* nodes within a circuit requires only the order of $t_v = d_v + r_v$, which a per-circuit constant cannot alter, so the binding unknown is r_v; *calibration* converts that order into a decision, and $L^\star$ is the only further quantity it needs. Section 5.2 shows that this decomposition, rather than integer exactness, is what the interventions probe.

Separately, the implemented descendant summary $c_{\mathrm{desc}}(v)$ is a capped child-sum recursion rather than $|\,\mathrm{desc}(v)|$, and in particular is not equivalent to r_v, so supplying it does not supply the second term of Equation 1; Appendix I gives the recursion.

2.3 Conditional information constraints

Both results, one stated here and one in Appendix A.2, are conditional, and we state the conditions precisely because the models we evaluate mostly violate them.

A *forward-only node-local* MPNN, one whose every aggregation is over parents alone and whose prediction at v reads $h_v^{(K)}$ only, computes a function of the K-hop ancestor neighborhood of v alone whenever its initial inputs are invariant to descendant-only modifications of G. Two DAGs differing only below v then yield identical representations while r_v and $\mathrm{CP}(v)$ differ, so no node-local readout is correct on both; Appendix A.2 states this precisely and proves it. Every hypothesis is load-bearing and none describes a model we train: the no-reverse models retain out-degree and descendant-count features, so their inputs are not descendant-free, and the default models augment each edge with

its reverse. The result therefore motivates the interventions of Section 5 rather than describing a baseline.

Proposition 1 (Constructive sufficiency)**.** *There exist parameters of an MPNN with separate parent and child channels, max aggregation, at least D propagation steps, and a broadcast of the graph-level maximum $L^\star$ that compute $\ell_\rightarrow(v)$, $\ell_\leftarrow(v)$, and $\mathrm{CP}(v)$ exactly, where $D = L^\star - 1$ is the maximum number of edges on a directed path in G.*

Proofs and the explicit recurrence are given in Appendix A. The construction is existential and does not assert that such parameters are reachable by gradient descent. The evaluated local models use at most eight propagation steps, whereas the depth-shift circuits have D between 50 and 522, so the proposition delimits what the depth sweeps of Section 5 do *not* test.

3 Models and Information Pathways

Default node inputs. Each node receives embeddings of masked gate type, gate family, and primary-wire index. The feature-enabled condition additionally supplies exact forward depth, in-degree, out-degree, sinusoidal sequence position, and the bucketed descendant count $b_v = \min\big(\lfloor \log_2(1+c_{\mathrm{desc}}(v)) \rfloor,\, 32\big)$, with c_{desc} as in Appendix I; the constant 32 bounds the bucket index and is not reached at the evaluated circuit sizes. These inputs are informative, so feature-enabled results measure use of supplied structure rather than recovery from adjacency.

Node-wise MLP and controlled information conditions. The Node-MLP applies eight residual feed-forward blocks independently to each node, at hidden width 192 and feed-forward width 768, closely matching GAT-L4 in parameters. Its encoder performs no message passing or node-to-node attention, and all six node-level heads operate only on node embeddings, so the pooled circuit representation the implementation exposes for interface compatibility never reaches the critical-path head.

We evaluate three information conditions: *default summaries*, the standard feature set; *graph-length broadcast*, which adds $L^\star = (\max_v d_v) + 1$ to every node; and *graph-length broadcast plus exact reverse depth*, which additionally supplies r_v computed from the DAG. Both scalars pass through a shared two-channel projection added to the node representation, with the reverse-depth channel fixed to zero in the broadcast-only condition, so the two intervention models have identical projection shape and are parameter matched at $3{,}007{,}948$ against $3{,}007{,}563$ for the default; at width 192 the projection adds exactly 385 parameters. Forward depth enters through the default structural encoding, so the three terms are supplied separately rather than as their residual, and Appendix C gives the normalization.

Local encoders. We evaluate GCN (Kipf & Welling, 2017), GAT (Veličković et al., 2018), and GIN (Xu et al., 2019). By default, each DAG edge is augmented with its reverse, allowing local communication from parents and children. The no-reverse condition retains only parent-to-child edges. These encoders use bidirectional neighborhoods but not relation-specific parent and child channels, receiving direction only through node features. This is deliberate: it is their standard deployed configuration, and channel separation, one of the four ingredients of Proposition 1, is instead tested where it is native, in the DAG-native encoder of Section 5.5.

Global encoders. The evaluated Perceiver (Jaegle et al., 2021; 2022) maps node features to a learned latent array through cross-attention and refines the latents through self-attention. Directed edge type and clipped relative depth enter through learned biases. Conclusions about this model concern the evaluated implementation and training configuration, rather than Perceiver architectures or global attention generally.

The graph transformer applies all-pairs attention directly to node representations, with directed edge-type and relative-depth biases in the full condition (Ying et al., 2021; Kreuzer et al., 2021). Its reduced-input condition is trained from initialization without depth, degree, descendant count, sequence position, or relative-depth bias, while retaining semantic inputs, wire identity, full attention, and parent/child edge-type bias.

Objective. All models use six node-level tasks: masked gate type, masked gate family, normalized depth regression, log descendant-count regression, critical-path classification with focal loss (Lin et al., 2017), and next-gate prediction, at weights 1.5, 0.01, 0.1, 0.1, 3.0, 0.8. Every architecture and condition shares these weights, the masking schedule, and a fixed zero-logit threshold; Appendix I gives the definitions, the depth-regression normalizer, and the global-encoder bias mechanisms.

4 Experimental Setup

Corpus. The training corpus contains 199,696 synthetic circuits generated from random, hardware-inspired, algorithmic, and topology-focused families. Circuits contain up to 1,024 gates and 128 qubits. Their longest-path lengths range from 4 to 546, with median 35, 95th percentile 157, and 99th percentile 236. We reserve 10% for validation.

Evaluation sets. The standard evaluation contains 116 MQT Bench circuits (Quetschlich et al., 2023) before exclusions. We report QFT-excluded aggregates as the primary standard-set metric because the separately encoded high-level QFT family is a distinct regime: its near-linear DAGs make every gate critical, its gate types are absent from the training vocabulary, and evaluated models therefore score exactly 0.000 or exactly 1.000 on it according to output bias rather than structural inference. Table 3 reports it separately and with QFT included.

The depth-shift evaluation contains 600 circuits before the same exclusion and 596 afterward, from the same generator family with a distinct seed and biased toward longer paths. QFT-excluded standard circuits have longest-path lengths from 2 to 86 with median 8, whereas depth-shift circuits range from 51 to 523 with median 99; no depth-shift circuit exceeds the observed training maximum, but 18.8% lie at or above the training 95th percentile, so this is a controlled upper-tail shift within observed support. The sets also differ in prevalence, with mean per-circuit critical fractions 0.492 and 0.335, so an all-positive predictor obtains QFT-excluded macro-F1 of 0.642 and 0.476; we therefore interpret F1 together with precision, recall, and threshold-free measures. Appendix B gives node counts and pooled fractions.

Training and reporting. Models use AdamW (Loshchilov & Hutter, 2019) at peak learning rate 5×10^{-4} with a OneCycle schedule (Smith & Topin, 2017), effective batch size 64, weight decay 0.01, gradient clipping at 1.0, and BF16 mixed precision (Micikevicius et al., 2018), for up to 60 epochs on a single H100 with early stopping on validation critical-path F1 at patience 12; it fired late or not at all for every reported configuration, so no result is prematurely truncated. Parameter matching holds within a band and does not equalize FLOPs, memory, or attention complexity (Appendix B). The primary metric is critical-path F1 computed per circuit and macro-averaged, with precision and recall macro-averaged independently, which prevents the substantially larger depth-shift circuits from dominating. All aggregates exclude QFT and uncertainties are standard deviations across all five seeds, where a population standard deviation is a coarse summary of training variability rather than a confidence interval; the substantially wider circuit-level bootstrap intervals appear in Table 2. With roughly thirty conditions and no multiple-comparison correction, marginal differences are descriptive and the load-bearing claims are those separated by many seed standard deviations or by intervals excluding zero.

5 Results

5.1 Holding architecture fixed isolates information access

The first three rows of Table 1 report the central controlled intervention. The default Node-MLP receives forward depth and local structural summaries but neither graph-level longest-path length nor exact reverse depth. Broadcasting $L^\star$ raises standard-set F1 from 0.573 ± 0.014 to 0.826 ± 0.008 for 385 parameters, from 3,007,563 to 3,007,948, which rules out capacity as the primary explanation of the baseline gap.

Because the deterministic rule reproduces every stored label exactly (Appendix C), residual error reflects imperfect realization rather than label ambiguity or missing information. The two increments differ in pattern: the broadcast gain contracts from $+0.253$ on the standard set to $+0.063$

under shift, whereas exact reverse depth adds 0.148 and 0.265, the graph-level statistic is primarily an in-distribution cue, whereas descendant-side information matters most under depth shift. (Appendix C).

5.2 Ranking versus calibration: the predicate is a threshold

It is natural to object that supplying two terms of a three-term identity trivializes the target. The objection is right in form but identifies the wrong binding term. The diagnostic it suggests fits a ridge regression predicting r_v from forward depth d_v, graph length $L^\star$, and gate count $|V|$, all available to the graph-length condition, and substitutes the estimate into Equation 1: it attains $R^2 = 0.845$ under depth shift yet collapses the otherwise exact rule to macro-F1 0.026, which invites the reading that integer exactness is binding. That reading is unsound, and the collapse is not a property of the equality *form*, since neither the threshold form nor any swept or per-circuit oracle offset recovers the rule (Appendix C.2).

The reason is an identity rather than a fit. Within a circuit $L^\star$ and $|V|$ are constants, so for any linear $\widehat{r}_v = \alpha d_v + \beta L^\star + \gamma |V| + \delta$ we have $d_v + \widehat{r}_v = (1+\alpha) d_v + c$ with c constant, and per-circuit ranking by the plug-in score is ranking by $\pm d_v$; we confirm this to machine precision on both sets (Table 5). High R^2 on r_v is moreover in tension with rank-relevance, since $r_v = (L^\star - 1) - d_v - \mathrm{slack}_v$, so a regressor holding d_v captures most of $\mathrm{Var}(r_v)$ through the $-d_v$ term while leaving the criticality signal slack_v unmodelled. Exact reverse depth therefore supplies a *direction* of information, within-circuit descendant-side variation provably absent from $(d_v, L^\star, |V|)$, rather than precision. Depth-shift macro AUPRC accordingly rises from 0.422 for the depth-equivalent linear score, to 0.512 for a linear probe on the full supplied feature set, to 0.511 for the *learned* node-wise encoder receiving those features, to 0.583 with the broadcast, and to 0.905 with exact r_v. That the learned encoder only matches a linear probe on its own inputs is the sharpest form of the claim: with these inputs, node-wise capacity buys no descendant-side information. Appendix C.4 recovers the ordering from the frozen representations. The identity constrains plug-in rules that hold the scoring function fixed, not two separately trained networks; Appendix A.1 records why a broadcast can still shift ranking metrics, and why its sign is not predicted.

The perturbation results are then predictions of this decomposition rather than separate observations. Constant bias on the supplied channel preserves the order of t_v and displaces only the threshold, whereas independent noise corrupts the order, so bias is less damaging at every magnitude; and because the predicate is monotone in t_v, the learned decision degrades gradually, crossing the no-reverse-depth reference at a fixed fraction of the median longest path on both sets (Appendices C.2 and C.3).

5.3 The same intervention transfers to a message-passing encoder

The intervention in Section 5.1 holds a node-wise architecture fixed. To test whether the effect is specific to the absence of message passing, we apply the identical graph-length broadcast to GAT-L4, using the same two-channel projection with the reverse-depth channel held at zero. The intervention adds exactly 385 parameters, from $3{,}017{,}547$ to $3{,}017{,}932$, matching the node-wise increment.

Paired per-seed differences are $+0.095 \pm 0.018$ on the standard set, positive on all five seeds, and $+0.011 \pm 0.015$ under depth shift, positive on four of five, so the standard-set gain is large relative to seed variation and the depth-shift gain comparable to it. On every seed the projection norm grows several-fold while its zero-input reverse-depth channel decays under weight decay alone, so the projection is active and supplies graph-level length only (Appendix F).

Two consequences follow. First, bidirectional message passing with mean-pooled readout does not by itself make graph-level longest-path length available to node representations. Second, the two sets differ in *what* the broadcast improves: in distribution macro AUPRC and AUROC move opposite to F1, from 0.978 to 0.966 and 0.987 to 0.979, so the gain is one of operating point, whereas under shift both improve, to 0.848 and 0.950. Section 5.2 predicts this, since $L^\star$ is a per-circuit constant that can only move the decision boundary: negligible at median longest path 8, material where $L^\star$ spans 51 to 523, and the cone-coverage measurement of Appendix F agrees. The standard-set decline is comparable to seed spread, so the supported claim is the absence of a ranking gain rather than its sign.

5.4 ARCHITECTURE RESULTS REFLECT DIFFERENT PATHWAYS

Table 1 compares representative feature-enabled models, measuring how architectures preserve and combine supplied information rather than recovery from connectivity.

Table 1: Critical-path F1 for the Node-MLP information conditions (top) and representative architectures (bottom), as QFT-excluded per-circuit means and population standard deviations over the stated seeds. "L" denotes layer count and "H" hidden width. The broadcast adds 385 parameters to both the Node-MLP and GAT-L4 and 553 to DAGNN; GIN-L8 is 5.07M and all other rows lie between 3.01M and 3.40M (Appendix B). The all-positive row is a prevalence-dependent floor rather than a model, and the default Node-MLP and the Perceiver fall below it on the standard and depth-shift sets respectively; the deterministic row is a label-integrity verification rather than a result.

Model or node-wise information	Seeds	Standard	Depth shift
Node-MLP, default summaries	5	$.573 \pm .014$	$.518 \pm .001$
Node-MLP, graph-length broadcast	5	$.826 \pm .008$	$.581 \pm .002$
Node-MLP, graph length + exact reverse depth	5	$.974 \pm .008$	$.846 \pm .007$
GAT-L4	5	$.814 \pm .020$	$.790 \pm .004$
GAT-L4, graph-length broadcast	5	$.909 \pm .008$	$.801 \pm .013$
GIN-L8	5	$.810 \pm .010$	$.814 \pm .008$
GAT-L8/H138, parameter-matched	5	$.817 \pm .007$	$.824 \pm .017$
DAGNN-L2, bidirectional (no $L^\star$)	5	$.791 \pm .002$	$.956 \pm .002$
DAGNN-L2, bidirectional + $L^\star$	5	$.883 \pm .002$	$\mathbf{.993 \pm .001}$
Perceiver	5	$.748 \pm .010$	$.408 \pm .011$
Graph transformer	5	$\mathbf{.981 \pm .008}$	$.982 \pm .002$
Graph transformer, reduced-input retraining	5	$.790 \pm .044$	$.720 \pm .006$
All-positive predictor (trivial reference)	—	.642	.476
Deterministic rule, exact d_v, r_v, $L^\star$	—	1.000	1.000

The default Node-MLP has standard-set precision 0.819 ± 0.015 but recall only 0.526 ± 0.012 (Table 7), so its F1 falls below the all-positive reference; its gate-type macro accuracy nonetheless reaches 0.409 ± 0.016, so the deficit is not a uniform failure of node-level learning. Circuit-level bootstrap intervals separate two regimes: on the standard set the broadcast and the local encoders are indistinguishable, with paired intervals spanning zero, whereas under shift GAT-L4 exceeds the broadcast by $+0.210$ and exact reverse depth exceeds GAT-L4 by $+0.056$, both excluding zero. The distinguishing quantity is descendant-side path information rather than the graph-level statistic alone, as Section 5.2 predicts, and the in-distribution parity is an artifact of the fixed threshold: macro AUPRC is 0.898 ± 0.006 against 0.978 ± 0.005, more than ten seed standard deviations apart. Table 2 and Appendix B.1 give all conditions and paired intervals.

5.5 A DAG-NATIVE ENCODER ACQUIRES DESCENDANT-SIDE INFORMATION WITHOUT A BROADCAST

The architectures above either receive descendant-side summaries as features or acquire them through finite local reach. A DAG-native encoder instead propagates along the partial order itself. We evaluate DAGNN (Thost & Chen, 2021) with two layers at width 276, chosen to match the GAT-L4 reference to within $+0.23\%$ at 3,024,602 parameters, in two conditions: bidirectional propagation alone, and the identical graph-length broadcast used in Section 5.3, which at this width adds $2H + 1 = 553$ parameters.

Bidirectional DAGNN reaches 0.956 ± 0.002 depth-shift F1 and macro AUPRC 0.975 with *no explicit* $L^\star$, exceeding the 0.846 and 0.905 of a node-wise model handed exact r_v *and* $L^\star$; adding the broadcast gives 0.993 ± 0.001 and 0.9985 ± 0.0004, near ceiling, where seed spread is not the operative variability, so we report the ordering rather than a separation. Three scope conditions bound the claim. The comparator is handicapped, since its supplied channel is scaled by $\max(1, L^\star - 1)$ and resolves the boundary only to $1/(L^\star - 1)$ (Appendix C.6). The arm varies full-diameter propagation, relation-specific channels, and GRU combination together, so it establishes sufficiency and not attribution. And length-bucketed batching makes the depth normalizer a per-circuit proxy for $L^\star$

that a depth-head diagnostic confirms the encoder holds, so "no $L^\star$" means no explicit broadcast; installing that channel in GAT-L4 lowers its depth-shift ranking (Appendix F).

On the standard set this condition reaches 0.883 against the graph transformer's 0.981, but median $L^\star$ there is 8 and depth binning localizes the entire deficit to the 38 circuits with $L^\star \leq 5$, which macro-averaging weights equally with the largest; from $L^\star \geq 6$ upward it matches or exceeds both GAT configurations. This is a three-of-four instantiation of Proposition 1, since DAGNN uses softmax-weighted sum where the construction requires max; Appendix F gives the cone-coverage pre-analysis, the batching control, and per-bin results.

5.6 Reverse communication and local reach

Removing reverse-edge augmentation while retaining structural summaries degrades local encoders by 0.120 to 0.130 paired F1 on the standard set and 0.150 to 0.190 under shift, so reverse communication provides information not recoverable from parent-to-child propagation alone, and grouped removal of the structural summaries is comparably broad. Additional local reach helps under shift, and the parameter-matched form of the claim is primary: GAT-L8 at width 138 raises depth-shift macro AUPRC from 0.827 to 0.863 at fixed capacity, retaining roughly 72% of the 0.127 F1 gain shown by the unmatched two-to-eight-layer sweep, which varies reach and capacity together; the standard set is null under both. Since eight steps remain far below the 50 to 522 that Proposition 1 requires, this is a sub-diameter effect rather than the sufficient-depth regime. Appendices D and E give per-encoder values, the unmatched comparison, and the matching caveat.

5.7 A latent-bottleneck encoder with reduced pairwise bias acquires little descendant-side signal.

The evaluated Perceiver reaches 0.748 ± 0.010 on the standard set but 0.408 ± 0.011 under depth shift, below the all-positive reference of 0.476 and with a circuit-level bootstrap interval excluding it. Threshold-free measurement attributes this to weak discrimination rather than a misplaced operating point, since retuning the threshold raises pooled depth-shift F1 only from 0.560 to 0.576: macro AUPRC is 0.494 against a macro prevalence of 0.335, the lowest of any condition we evaluate, below the 0.545 of a linear probe on within-circuit-varying supplied features and barely above the 0.422 of forward depth alone. This configuration therefore acquires essentially no rank-relevant descendant information beyond a linear function of its inputs, its node ordering is worse than that of every local encoder, and depth-shift F1 stays between 0.39 and 0.41 as latent count varies from 32 to 512. Two scope conditions bound the claim. The Perceiver receives pairwise structural bias only as a per-node scalar reduced from the $N \times N$ tensor the graph transformer uses pair-resolved, so the two are not matched on that channel and the gap cannot separate a latent-routing limitation from the weaker bias (Appendix I); and four of five latent-count entries are single runs. The supported claim is configuration-specific: latent attention as instantiated here does not by itself yield graph-relative reasoning on this predicate.

The full graph transformer is a counterexample to any universal local-versus-global ranking, reaching 0.981 ± 0.008 and 0.982 ± 0.002 with structural summaries and graph-relative attention biases. Retraining without node summaries or relative-depth bias, but retaining semantic inputs, wire identity, all-pairs attention and directed edge-type bias, yields 0.790 ± 0.044 and 0.720 ± 0.006: full attention recovers substantial structural signal, while supplied structure accounts for much of the full-condition performance.

A third condition, removing the structural pathway only at inference, degrades bimodally and is a feature-dependence diagnostic rather than a matched comparison (Appendix H). The utility check of Appendix J lies outside the predicate.

6 Related Work

Local GNNs aggregate over explicit neighborhoods (Kipf & Welling, 2017; Veličković et al., 2018; Xu et al., 2019), so what they distinguish depends on the aggregation rule and on the portion of the graph reachable within finite depth, and long-range work studies the resulting limitations including oversquashing (Alon & Yahav, 2021; Topping et al., 2022; Dwivedi et al., 2022). Graph transformers

combine global attention with structural encodings or biases (Ying et al., 2021; Kreuzer et al., 2021), DAG-specific architectures propagate in topological order (Thost & Chen, 2021) or restrict attention by reachability (Luo et al., 2023), and Perceiver models route through latent arrays (Jaegle et al., 2021; 2022). Expressiveness results relate message passing to depth and width (Loukas, 2020), and neural algorithmic reasoning studies learned execution of procedures (Veličković & Blundell, 2021).

Our contribution is not the established observation that structural encodings or bidirectional communication help. For a DAG predicate with an exact dependency decomposition, we instead identify the specific ancestor-side, descendant-side, and graph-level quantities that must reach a node-level predictor, use parameter-matched information interventions to separate access to those quantities from node-processing capacity, and exploit the decomposition for direct information-sufficiency checks. The framework also makes a falsifiable prediction about a comparator we do not evaluate. Reachability-restricted DAG attention (Luo et al., 2023) has receptive field exactly $\mathrm{cone}(v)$, which is the object the cone analysis of Appendix F characterizes *a priori* and independently of any model. Such an architecture should therefore acquire r_v without difficulty but should require an explicit $L^\star$ broadcast in proportion to the measured cone deficit: more on shallow wide circuits, where $L^\star$ is absent from the descendant-side cone at 45.0% of non-critical standard-set nodes, than on deep ones, where the figure is 19.2%. Unrestricted all-pairs attention, by contrast, can compute $\max_u d_u$ in a single layer and so should need no broadcast at all, which is what we observe in Appendix H. Separately, multi-task graph pretraining is well studied in molecular representation learning (Hu et al., 2020; You et al., 2021); our objective combines masked semantic prediction with deterministic DAG labels. The routing check relates to learned compiler-cost models (Mendis et al., 2019; Adams et al., 2019), quantum routing (Li et al., 2019; Javadi-Abhari et al., 2024), and heterogeneous quantum scheduling (Ravi et al., 2022).

7 Limitations and Conclusion

The experiments concern quantum circuit DAGs, one structural predicate, and one controlled same-generator upper-tail depth shift. Circuit DAGs have degree bounded by gate arity, qubit-parallel branches and characteristic reconvergence, so transfer to irregular DAGs such as abstract syntax trees is untested, as is any predicate lacking an exact additive decomposition. The Perceiver and reduced-input graph-transformer results each concern one implementation, the DAG-native arm supports sufficiency and not attribution (Appendix I), and exact reverse depth is graph-derived, so that Node-MLP condition diagnoses information sufficiency rather than providing a practical encoder. The rank-equivalence analysis concerns linear estimators and per-circuit ranking; pooled metrics compare across circuits, where $L^\star$ does enter. The most serious limitation is that every depth-regime claim rests on same-generator synthetic circuits: the only external benchmark comprises 103 circuits with median $L^\star = 8$, where four hops already span most graphs, and the depth-shift set shares its generator, features and gate vocabulary with the training corpus, so it tests extrapolation in path length but not transfer across sources. We also omit reachability-restricted DAG attention (Luo et al., 2023), for which Section 6 states our cone analysis's prediction, and a minimal two-pass topological encoder. Three runs would falsify rather than extend specific claims: an $L^\star$-free reverse-depth condition, a globally fixed depth normalizer for the DAG-native arm, and an $L^\star$-independent encoding of r_v (Appendices F and C.6). The cheapest test of generality is a *weighted* critical path, where $s_v \leq L^\star$ and the ranking–calibration split survive but the additive decomposition into stored features does not.

Within this scope the evidence supports a consistent conclusion. The same node-wise architecture moves from 0.573 to 0.826 to 0.974 standard-set F1 and from 0.518 to 0.581 to 0.846 under shift as graph-level and then descendant-side path information are added. Local receptive field, reverse communication, topological propagation, latent routing, and graph-aware global attention can all be understood as mechanisms for acquiring or combining these quantities. Successful structural prediction is therefore governed not by architecture labels alone, but by whether the representation obtains reliable access to the information the target demands.

Reproducibility Statement

Sections 3 and 4 specify input conditions, objectives, datasets, metrics, optimization, exclusions, and seed reporting. Appendices A through J give the proof details, the deterministic feature construction and exact reverse-depth recurrence, the aggregation and exclusion rules, every ablation condition, the receptive-field and depth-tail controls, the graph-transformer conditions, the objective and masking definitions, and the routing-probe protocol in sufficient detail to reimplement every reported condition.

Ethics Statement

This work uses synthetic circuits and public benchmark quantum circuits and involves no human participants, personal data, or sensitive attributes. The MQT Bench circuits are used under their public open-source license. The primary costs are computational and environmental; experiments use relatively small single-GPU models.

AI Use Statement

Generative AI tools assisted with manuscript organization and prose revision. They were not used to generate experimental measurements, datasets, model checkpoints, or unverified proofs. The authors checked all mathematical arguments, numerical values, citations, and experimental descriptions and take responsibility for the final manuscript.

References

Andrew Adams, Karima Ma, Luke Anderson, Riyadh Baghdadi, Tzu-Mao Li, Michaël Gharbi, Benoit Steiner, Steven Johnson, Kayvon Fatahalian, Frédo Durand, and Jonathan Ragan-Kelley. Learning to optimize halide with tree search and random programs. *ACM Transactions on Graphics (TOG)*, 38(4):121:1–121:12, 2019.

Uri Alon and Eran Yahav. On the bottleneck of graph neural networks and its practical implications. In *International Conference on Learning Representations (ICLR)*, 2021.

Vijay Prakash Dwivedi, Ladislav Rampášek, Mikhail Galkin, Ali Parviz, Guy Wolf, Anh Tuan Luu, and Dominique Beaini. Long range graph benchmark. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2022.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay Pande, and Jure Leskovec. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2020.

Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. In *International Conference on Machine Learning (ICML)*, 2021.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver IO: A general architecture for structured inputs and outputs. In *International Conference on Learning Representations (ICLR)*, 2022.

Ali Javadi-Abhari, Matthew Treinish, Kevin Krsulich, Christopher J. Wood, Jake Lishman, Julien Gacon, Simon Martiel, Paul D. Nation, Lev S. Bishop, Andrew W. Cross, Blake R. Johnson, and Jay M. Gambetta. Quantum computing with Qiskit. *arXiv preprint arXiv:2405.08810*, 2024.

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

Devin Kreuzer, Dominique Beaini, William L. Hamilton, Vincent Létourneau, and Prudencio Tossou. Rethinking graph transformers with spectral attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Gushu Li, Yufei Ding, and Yuan Xie. Tackling the qubit mapping problem for NISQ-era quantum devices. In *International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2019.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Andreas Loukas. What graph neural networks cannot learn: depth vs width. In *International Conference on Learning Representations (ICLR)*, 2020.

Yuankai Luo, Veronika Thost, and Lei Shi. Transformers over directed acyclic graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Charith Mendis, Alex Renda, Saman Amarasinghe, and Michael Carbin. Ithemal: Accurate, portable and fast basic block throughput estimation using deep neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *International Conference on Learning Representations (ICLR)*, 2018.

Nils Quetschlich, Lukas Burgholzer, and Robert Wille. MQT bench: Benchmarking software and design automation tools for quantum computing. *Quantum*, 7:1062, 2023.

Gokul Subramanian Ravi, Kaitlin N. Smith, Prakash Gokhale, and Frederic T. Chong. Quantum computing in the cloud: Analyzing job and machine characteristics. In *IEEE International Symposium on Workload Characterization (IISWC)*, 2022.

Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. *arXiv preprint arXiv:1708.07120*, 2017.

Veronika Thost and Jie Chen. Directed acyclic graph neural networks. In *International Conference on Learning Representations (ICLR)*, 2021.

Jake Topping, Francesco Di Giovanni, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M. Bronstein. Understanding over-squashing and bottlenecks on graphs via curvature. In *International Conference on Learning Representations (ICLR)*, 2022.

Petar Veličković and Charles Blundell. Neural algorithmic reasoning. *Patterns*, 2(7), 2021.

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

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations (ICLR)*, 2019.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. Do transformers really perform bad for graph representation? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yuning You, Tianlong Chen, Yang Shen, and Zhangyang Wang. Graph contrastive learning automated. In *International Conference on Machine Learning (ICML)*, 2021.

A Proof Details

A.1 The threshold form

Proposition 2 (Longest path through a node)**.** *For every $v \in V$ the node count of a longest directed path through v is exactly $s_v = d_v + r_v + 1$. Consequently $s_v \le L^\star$, with equality if and only if* $\mathrm{CP}(v) = 1$.

Proof. Let $\pi_\rightarrow$ be a longest directed path ending at v and $\pi_\leftarrow$ a longest directed path starting at v. Since G is acyclic the two meet only at v, so their concatenation is a directed path through v on $\ell_\rightarrow(v) + \ell_\leftarrow(v) - 1 = s_v$ nodes; a path of that length therefore exists. Conversely, any directed path through v splits at v into a path ending at v and a path starting at v, neither longer than the corresponding maximum, so no path through v exceeds s_v. Hence s_v is both attained and maximal. Taking the maximum over V gives $s_v \le L^\star$, and $s_v = L^\star$ holds exactly when some longest directed path of G passes through v, which is the definition of $\mathrm{CP}(v) = 1$. □

Two corollaries are used in the main text.

One-sidedness. Because $s_v > L^\star$ is impossible, $\mathbb{I}[s_v = L^\star]$ and $\mathbb{I}[s_v \ge L^\star]$ agree on every DAG. The predicate is therefore a one-sided threshold in $t_v = d_v + r_v$, and with $\mathrm{slack}_v = L^\star - s_v \ge 0$ we have $\mathrm{CP}(v) = \mathbb{I}[\mathrm{slack}_v = 0]$. This is why the failure of the plug-in rules in Appendix C.2 is attributable to the estimate rather than to the equality form: the threshold form is available and does no better.

Rank invariance, and its scope. $L^\star$ is constant within a circuit, so any strictly increasing function of t_v induces the same within-circuit ordering of nodes by criticality as t_v itself. No per-circuit ranking metric can therefore depend on $L^\star$ *when the scoring function is held fixed*, and the rank-equivalence identity of Section 5.2 is the specialization of this to affine estimators of r_v. The qualification is necessary and is not a technicality. Two networks trained with and without an $L^\star$ input do not share a scoring function: supplying the scalar changes the shared trunk that all six heads read, so the broadcast may alter node ranking as a representation-learning effect, and the corollary makes no prediction about the sign. Table 4 shows both signs, with macro AUPRC rising for the Node-MLP conditions and falling in distribution for GAT-L4, which is consistent with a retraining effect and inconsistent with a fixed-score recalibration.

A.2 Descendant blindness

Definition 1 (Forward-only node-local MPNN)**.** *For $k = 1, \dots, K$ let*

$$h_v^{(k)} = \phi^{(k)}\left(h_v^{(k-1)}, \bigoplus_{u \in \mathrm{pa}(v)} \psi^{(k)}(h_u^{(k-1)})\right).$$

Such a network is forward-only*, since every aggregation is over parents alone, and* node-local *if its prediction at v is a function of $h_v^{(K)}$ only, receiving no separately computed unrestricted graph representation. Its initial inputs $h_v^{(0)}$ are* descendant-free *if they are invariant to every modification of G that alters only* $\mathrm{desc}(v)$.

Proposition 3 (Descendant blindness)**.** *Let f be a forward-only node-local MPNN of depth K with descendant-free initial inputs. Then $h_v^{(K)}$ is a function of the K-hop ancestor neighborhood of v alone. Consequently there exist pairs of DAGs, differing only in the descendants of v, on which $h_v^{(K)}$ is identical while r_v and $\mathrm{CP}(v)$ differ, so no node-local readout of f can be correct on both.*

Let $\mathrm{anc}_{\le k}(v)$ denote nodes that can reach v through at most k directed edges. We prove by induction that $h_v^{(k)}$ is a function only of initial representations in

$$\{v\} \cup \mathrm{anc}_{\le k}(v).$$

At $k = 0$, the statement follows by definition. Suppose it holds at layer $k-1$. The layer-k update at v receives $h_v^{(k-1)}$ and representations of parents $u \in \mathrm{pa}(v)$. By induction, each parent representation

depends only on

$$\{u\} \cup \mathrm{anc}_{\le k-1}(u),$$

which lies in $\{v\} \cup \mathrm{anc}_{\le k}(v)$. To close the argument it remains to note that the initial inputs at those nodes are themselves unchanged by a descendant-only modification. If $u \in \mathrm{anc}_{\le k}(v)$ then $\mathrm{desc}(v) \subseteq \mathrm{desc}(u)$, since any directed path from v extends a directed path from u through v; so a modification confined to $\mathrm{desc}(v)$ is also confined to $\mathrm{desc}(u)$, and descendant-free inputs at u are invariant to it by Definition 1. Every term determining $h_v^{(K)}$ is therefore fixed, and descendants never enter the representation.

For a critical-path witness, construct two DAGs with the same ancestor chain ending at v and a competing path of fixed length P. In the first, the descendant chain below v is short, so v is not on a longest path. Extend only that descendant chain until the path through v exceeds P. The label changes while the forward-only representation at v remains fixed. No node-local readout can classify both graphs correctly.

The same argument applies to any target changed by a descendant-only modification. It does not apply if descendant-derived quantities are supplied initially, if messages propagate from children to parents, or if predictions receive unrestricted graph-level context.

A.3 Constructive sufficiency

Maintain scalar channels $f_v^{(k)}$ and $g_v^{(k)}$, initialized to one:

$$f_v^{(k+1)} = \max\left(f_v^{(k)}, 1 + \max_{u \in \mathrm{pa}(v)} f_u^{(k)}\right),$$

$$g_v^{(k+1)} = \max\left(g_v^{(k)}, 1 + \max_{w \in \mathrm{ch}(v)} g_w^{(k)}\right),$$

where the inner term is dropped when the corresponding index set is empty. After $D = L^\star - 1$ iterations, where D is the maximum number of edges on a directed path,

$$f_v^{(D)} = \ell_{\to}(v), \qquad g_v^{(D)} = \ell_{\leftarrow}(v).$$

A graph-level maximum gives

$$L^\star = \max_{u \in V} f_u^{(D)}.$$

Broadcasting $L^\star$ permits the comparison

$$f_v^{(D)} + g_v^{(D)} - 1 = L^\star.$$

This is an existence construction. It does not assert that finite-width GAT or GIN layers trained by gradient descent will implement the recurrence.

B Evaluation and Aggregation Details

The standard set contains 116 circuits before exclusion and 103 after excluding the separately encoded QFT family. The depth-shift set contains 600 circuits before exclusion and 596 afterward.

Critical-path F1, precision, recall, gate-type macro accuracy, and depth correlation are first computed per circuit. Each seed-level aggregate is the arithmetic mean over retained circuits. Reported uncertainty is the population standard deviation over seed-level aggregates. Because F1, precision, and recall are averaged independently over circuits, aggregate F1 need not equal the harmonic mean of aggregate precision and recall.

The QFT-excluded standard set contains 5,985 nodes and has pooled critical-node fraction 0.486. The depth-shift set contains 367,290 nodes and has pooled critical-node fraction 0.305. Per-circuit macro-averaging prevents large depth-shift circuits from dominating reported aggregates. Parameter matching holds within a band rather than universally. The node-wise MLP, the three four-layer local encoders, and the Perceiver span 3.01M–3.30M parameters, a spread of approximately 10%, and the two graph-transformer conditions are 3.32M and 3.40M. The layer sweeps are deliberately not

Table 2: Circuit-level bootstrap intervals on the QFT-excluded standard set ($n = 103$ circuits, $10{,}000$ resamples over circuits with a fixed resampling seed, applied to per-circuit F1 averaged over training seeds). Seed-level standard deviations elsewhere in the paper describe training variability only and are substantially narrower than circuit-level sampling variability. Paired standard-set differences discussed in Section 5.4 are $+0.012$ $[-0.038, +0.067]$ for the broadcast against GAT-L4 and $+0.016$ $[-0.029, +0.065]$ against GIN-L8; under depth shift GAT-L4 exceeds the broadcast by $+0.210$ $[+0.201, +0.218]$, GIN-L8 by $+0.233$ $[+0.225, +0.241]$, and reverse depth exceeds GAT-L4 by $+0.056$ $[+0.040, +0.071]$.

Condition	Mean	95% CI
Node-MLP, default summaries	.573	[.514, .632]
Node-MLP, graph-length broadcast	.826	[.794, .856]
Node-MLP, graph length + reverse depth	.974	[.961, .984]
GAT-L4	.814	[.749, .876]
GIN-L8	.810	[.750, .865]
Perceiver	.748	[.714, .781]
Graph transformer	.981	[.973, .988]

Table 3: Sensitivity to the separately encoded QFT family on the standard set. "All" includes all 116 circuits; "No QFT" excludes 13 QFT circuits; "QFT only" reports the 13-circuit family. Values are per-circuit macro F1 and population standard deviations over five seeds.

Condition	All	No QFT	QFT only
Node-MLP, default summaries	$.509 \pm .012$	$.573 \pm .014$	$.000 \pm .000$
Node-MLP, graph-length broadcast	$.733 \pm .007$	$.826 \pm .008$	$.000 \pm .000$
Node-MLP, graph-length broadcast + exact reverse depth	$.864 \pm .007$	$.974 \pm .008$	$.000 \pm .000$
GAT-L4	$.722 \pm .016$	$.814 \pm .020$	$.000 \pm .000$
Perceiver	$.776 \pm .007$	$.748 \pm .010$	$1.000 \pm .000$
Graph transformer	$.983 \pm .007$	$.981 \pm .008$	$1.000 \pm .000$

parameter-matched across depth, so Table 12 varies receptive field and capacity together and GIN-L8 in Table 1 is 5.07M rather than approximately 3M. Parameter matching also does not equalize FLOPs, memory, or attention complexity.

The QFT family is degenerate in a specific and diagnosable way. Its label positive rate is exactly 1.000: every gate lies on a longest path, so an all-positive predictor scores 1.000. Gate-type macro accuracy on this family is 0.000 for every evaluated model, including the two that reach critical-path F1 of 1.000, which is consistent with the separately encoded high-level QFT gate types being absent from the training vocabulary. Conditional on this semantic blindness, critical-path F1 on QFT is determined by each model's default output bias rather than by structural inference: the Node-MLP conditions and GAT-L4 score 0.000 with precision also 0.000, indicating that they emit no positive predictions there, while the Perceiver and the graph transformer score exactly 1.000 with precision and recall also exactly 1.000. The perfect scores are therefore not evidence of structural reasoning on this family, and the zero scores are not evidence of a critical-path-specific failure. Reporting the family separately in Table 3 avoids attributing either outcome to representational capability.

B.1 Threshold-free evaluation

Because critical-path prevalence differs between the two evaluation sets, F1 at a fixed operating point conflates discriminative ability with calibration. We therefore recompute all headline conditions using threshold-free measures. Node-level critical-path logits are extracted from each best-validation-F1 checkpoint without gate-type masking, and we report per-circuit macro-averaged AUPRC and AUROC alongside pooled F1 at the fixed zero-logit threshold and pooled F1 at the best achievable threshold. The last two quantities are both pooled over nodes, so their difference isolates the cost of the fixed operating point.

Four conclusions follow from Table 4.

Table 4: Threshold-free evaluation of all headline conditions. AUPRC and AUROC are macro-averaged over circuits; the two F1 columns are pooled over nodes, so that their difference measures calibration alone. Uncertainties are population standard deviations over seeds. Because AUPRC is computed per circuit and then averaged, the correct random-classifier reference is the mean per-circuit prevalence, 0.492 on the standard set and 0.335 under depth shift, rather than the pooled prevalence of 0.486 and 0.305 that governs the pooled F1 columns. Macro F1 recomputed by this harness, which is not shown here, agrees with the values in Table 1 to within 0.011 for every condition but one; the reduced-input graph transformer differs by 0.029 on the standard set, within its seed spread. Thresholds are swept on a finite grid and all entries are rounded to three decimals, so Δ can be marginally negative, as for the depth-shift graph transformer.

Set	Condition	AUPRC	AUROC	$F1_0$	$F1_{best}$	Δ
Standard	Node-MLP, default	$.832 \pm .005$	.859	.728	.753	$+.025$
	Node-MLP, $L^\star$	$.898 \pm .006$	.923	.799	.807	$+.008$
	Node-MLP, $L^\star + r_v$	$.997 \pm .001$	.999	.958	.961	$+.003$
	GAT-L4	$.978 \pm .005$	.987	.881	.882	$+.001$
	GAT-L4, $L^\star$	$.966 \pm .014$	.979	.900	.905	$+.005$
	GIN-L8	$.929 \pm .010$	.938	.862	.869	$+.007$
	Perceiver	$.852 \pm .003$	.800	.759	.823	$+.064$
	Graph transformer	$.998 \pm .001$	.999	.969	.974	$+.005$
	Graph transformer, reduced	$.815 \pm .055$	.801	.751	.763	$+.012$
	GAT-L8/H138, matched	$.973 \pm .010$	.982	.882	.884	$+.002$
	DAGNN, bidir	$.970 \pm .005$	.968	.948	.952	$+.004$
	DAGNN, bidir + $L^\star$	$.978 \pm .006$	.982	.969	.969	$+.000$
Depth shift	Node-MLP, default	$.511 \pm .002$	.695	.538	.539	$+.001$
	Node-MLP, $L^\star$	$.583 \pm .002$	.768	.604	.605	$+.001$
	Node-MLP, $L^\star + r_v$	$.905 \pm .012$	.955	.824	.828	$+.004$
	GAT-L4	$.827 \pm .005$	.940	.826	.833	$+.007$
	GAT-L4, $L^\star$	$.848 \pm .016$	.950	.840	.840	$+.000$
	GIN-L8	$.851 \pm .005$	.950	.848	.852	$+.004$
	Perceiver	$.494 \pm .002$	.649	.560	.576	$+.016$
	Graph transformer	$.995 \pm .001$	.999	.984	.983	$-.001$
	Graph transformer, reduced	$.762 \pm .003$	.881	.774	.776	$+.002$
	GAT-L8/H138, matched	$.863 \pm .016$	.956	.853	.859	$+.006$
	DAGNN, bidir	$.975 \pm .000$	.995	.966	.966	$+.000$
	DAGNN, bidir + $L^\star$	$.999 \pm .000$	1.000	.993	.991	$-.002$

First, the fixed zero-logit threshold is close to optimal for nearly every condition. Across both sets the gap between fixed-threshold and best-threshold pooled F1 is at most 0.008 under depth shift and at most 0.025 on the standard set, for every condition but one. The reported comparisons are therefore not artifacts of a shared operating point, and tuning the threshold per model would not materially change any ordering. The single exception is the Perceiver on the standard set, where the gap is 0.064.

Second, the information ladder is threshold-independent. Macro AUPRC rises from 0.832 to 0.898 to 0.997 on the standard set and from 0.511 to 0.583 to 0.905 under depth shift as graph-level and descendant-side path quantities are added, reproducing the ordering and approximate magnitudes of Table 1 without reference to any threshold.

Third, AUPRC separates conditions that F1 does not. On the standard set, the graph-length broadcast and GAT-L4 differ by 0.080 in macro AUPRC, a gap of more than ten seed standard deviations, while their macro F1 differs by 0.012 under the primary harness of Table 1 and by 0.008 when recomputed here without gate-type masking; the two harnesses agree to within the tolerance stated in the caption, and the two values should not be mixed. Section 5.4 discusses this case.

Fourth, applying the broadcast to GAT-L4 improves calibration and node ranking asymmetrically across the two sets. On the standard set it raises macro F1 from 0.814 to 0.909, while macro AUPRC falls from 0.978 to 0.966 and macro AUROC from 0.987 to 0.979; both threshold-free measures therefore move opposite to F1, and the gain is one of operating point rather than of node ordering. Under depth shift all three improve, from 0.790 to 0.801 in macro F1, from 0.827 to 0.848

in AUPRC, and from 0.940 to 0.950 in AUROC. Supplying a graph-level statistic therefore need not improve ranking and calibration together, which reinforces the separation between the two that Section 5.4 draws for the node-wise conditions. The standard-set AUPRC change is of comparable magnitude to the intervention condition's seed standard deviation, so we treat the absence of a ranking gain, rather than the sign of the decline, as the supported claim.

C Controlled Node-MLP Information Conditions

C.1 Deterministic feature construction

The deterministic classifier

$$\widehat{\mathrm{CP}}(v) = \mathbb{I}\left[d_v + r_v = L^\star - 1\right]$$

achieves F1, precision, and recall equal to 1.0 on every stored standard and depth-shift evaluation circuit, as does the equivalent threshold form. Consequently, the residual error of the learned graph-length broadcast plus exact reverse-depth model reflects neither label uncertainty nor insufficiency of the supplied path quantities; Appendix C.6 attributes it instead to the resolution of the normalized reverse-depth channel.

The stored data contain forward depth, descendant count, ancestor count, direct predecessor and successor counts, critical-path labels, full edge indices, critical-path length, and maximum depth. Exact reverse longest-path depth is computed by reverse-topological dynamic programming.

For every evaluation circuit, the resulting reverse depths satisfy the edge invariant

$$r_u \geq r_v + 1 \qquad \text{for every edge } u \to v,$$

and

$$\mathrm{CP}(v) = \mathbb{I}\left[d_v + r_v = L^\star - 1\right]$$

reproduces every stored critical-path label exactly. The graph-length broadcast plus exact reverse-depth condition supplies these deterministic quantities as node features. Conditional on them, the learned Node-MLP uses no message passing, node-to-node attention, or adjacency-based computation.

The graph-length broadcast and graph-length-plus-reverse-depth conditions use the same two-channel intervention projection. For the broadcast, $L^\star$ is scaled by the configured maximum depth and copied to every valid node; in the exact reverse-depth condition, r_v is scaled by $\max(1,\, L^\star - 1)$, yielding a relative position along a longest-path-length scale. The reverse-depth channel is zero in the graph-length-only condition. Consequently, both have exactly $3{,}007{,}948$ trainable parameters.

The intervention does not provide the equality residual itself. The learned model receives separately normalized graph-length and reverse-depth channels and must still combine them with forward depth through its node-wise nonlinear predictor.

C.2 Rule form, rank equivalence, and slack-space measurement

For each evaluation set we fit a ridge regression ($\alpha = 1$, standardized features) predicting exact reverse depth r_v from three quantities the graph-length broadcast condition already receives: forward depth d_v, graph-level longest-path length $L^\star$, and gate count $|V|$. The model is fit and evaluated on the same QFT-excluded circuits, so its accuracy is an upper bound. We substitute the estimate into the decision rule in both the equality and the equivalent threshold form, and separately measure what the estimate contributes to node ranking.

Section 5.2 states the consequences; five details are recorded here.

The two error modes of the equality form are strongly asymmetric. A one-node underestimate of r_v falsely rejects every critical node, 48.6% and 30.5% of nodes on the two sets, whereas a one-node overestimate falsely accepts only the 9.4% and 10.2% at $\mathrm{slack}_v = 1$, and the 17.1% and 32.5% of nodes at $\mathrm{slack}_v \geq 8$ are unaffected by either. The threshold form adjudicates none of the latter.

The plug-in result is computed after rounding $\widehat{r}_v$ to the nearest integer, and no rounding convention recovers the rule. That failure is not attributable to the equality form: the threshold form is no

Table 5: What a linear estimate of reverse depth does and does not supply. The plug-in rows replace exact r_v with a ridge estimate from d_v, $L^\star$, and $|V|$, under both rule forms and under swept thresholds; the ranking rows use no threshold at all. Learned rows are repeated from Table 1 for comparison. All values exclude QFT.

Quantity	Standard	Depth shift
R^2 of ridge estimate $\widehat{r}_v$ (pooled)	0.689	0.845
R^2 restricted to within-circuit variation, target r_v	0.417	0.839
R^2 restricted to within-circuit variation, target slack_v	0.253	0.018
Mean absolute error of $\widehat{r}_v$ (edges)	5.79	13.90
Fraction of nodes with $\lfloor \widehat{r}_v \rceil = r_v$	0.070	0.026
Exact r_v, equality form	1.000	1.000
Exact r_v, threshold form	1.000	1.000
Plug-in $\widehat{r}_v$, equality form	0.078	0.026
Plug-in $\widehat{r}_v$, threshold form at $\tau = 0$	0.083	0.019
Plug-in, threshold form, global oracle τ	0.655	0.476
Plug-in, threshold form, per-circuit oracle τ	0.798	0.502
All-positive reference	0.642	0.476
Macro AUPRC of exact $d_v + r_v$ (no $L^\star$)	1.000	1.000
Macro AUPRC of $d_v + \widehat{r}_v$	0.847	0.422
Macro AUPRC of d_v alone	0.847	0.422
Learned model, graph-length broadcast (no r_v)	0.826	0.581
Learned model, graph length + exact r_v	0.974	0.846

better, and sweeping the offset reaches the all-positive reference under depth shift and stays below the 0.826 that the same learned architecture attains with no reverse-depth input at all.

The equality of the three macro-AUPRC rows is exact rather than approximate. Within a circuit $L^\star$ and $|V|$ are constants, so $d_v + \widehat{r}_v$ is affine in d_v; per-circuit ranking is invariant to affine maps up to sign, and the fitted raw coefficient on d_v is -0.579 on the standard set and -0.865 under shift, both above -1, so the sign is preserved. Measured agreement with ranking by d_v is $0.00\mathrm{e}{+00}$ in both AUPRC and AUROC on both sets.

The within-circuit variance of r_v decomposes as $\mathrm{Var}_w(d_v) + \mathrm{Var}_w(\mathrm{slack}_v) + 2\,\mathrm{Cov}_w$, with the three terms contributing 0.992, 0.164 and -0.156 of the total under depth shift and 1.189, 0.780 and -0.969 on the standard set. Per-circuit ordinary-least-squares slopes of r_v on d_v have median -0.967 under shift with 77% inside $[-1.1, -0.9]$, against median -0.333 and 14% on the standard set, so the near-deterministic $r_v \approx \text{const} - d_v$ relation that inflates R^2 while cancelling in the sum is a depth-shift phenomenon specifically.

Finally, the same probe measured in slack space rather than on r_v orders the feature sets as their downstream AUPRC does. On the full supplied feature set, adding in-degree, out-degree and the bucketed descendant count, within-circuit R^2 on slack rises from 0.018 to 0.127 under shift and macro AUPRC of $d_v + \widehat{r}_v$ from 0.422 to 0.512; dropping the circuit constants and predicting slack directly from the within-varying features reaches 0.545. The learned node-wise encoder with default summaries attains 0.511, statistically indistinguishable from the 0.512 of a linear probe on the same inputs.

C.3 Error tolerance of the learned predictor

Appendix C.2 shows that no plug-in rule form recovers the predicate from a linear estimate of reverse depth, because that estimate carries no within-circuit descendant signal. We now hold the supplied channel exact but perturb it, to measure how much error the *learned* predictor absorbs when the signal is present. The five trained graph-length-plus-reverse-depth checkpoints are re-evaluated with no retraining.

Reverse depth is computed exactly, the label-reconstruction identity is verified on the unperturbed tensor, and only then is a perturbation added before the channel is normalized by $\max(1, L^\star - 1)$. We use two error models. Additive noise draws an independent integer uniformly from $\{-\varepsilon, \ldots, +\varepsilon\}$ at

each node, which resembles estimator error. Constant bias adds $+\varepsilon$ to every node, which preserves the ranking of $d_v + r_v$ exactly while displacing the equality. Perturbed values are clipped below at zero, so nodes near sinks receive slightly less effective noise than the nominal ε.

Table 6: Critical-path F1 under perturbation of the supplied reverse-depth channel. Values are QFT-excluded per-circuit means and population standard deviations over the same five seeds as Table 1. Reference rows are the learned graph-length-only condition, which receives no reverse depth at all, and the plug-in exact rule from Table 5. At $\varepsilon = 0$ the two error models coincide, so a single value is reported per set.

	Standard		Depth shift	
Perturbation	Noise	Bias	Noise	Bias
$\varepsilon = 0$ (exact)	$.974 \pm .008$		$.846 \pm .007$	
$\varepsilon = 1$	$.813 \pm .012$	$.885 \pm .003$	$.790 \pm .014$	$.804 \pm .011$
$\varepsilon = 2$	$.756 \pm .009$	$.838 \pm .002$	$.732 \pm .016$	$.753 \pm .011$
$\varepsilon = 4$	$.696 \pm .012$	$.760 \pm .001$	$.657 \pm .014$	$.690 \pm .009$
$\varepsilon = 8$	$.644 \pm .017$	$.713 \pm .001$	$.581 \pm .010$	$.614 \pm .004$
No reverse depth supplied	$.826 \pm .008$		$.581 \pm .002$	
Exact rule, plug-in estimate	.078		.026	

Section 5.2 states the conclusions. Because perturbed values are clipped below at zero, the reported degradation is a mild underestimate.

C.4 What frozen representations linearly encode about slack

Appendix C.2 shows that R^2 on r_v does not measure rank-relevant descendant information. Probing frozen representations for r_v, as we did initially, inherits that defect and additionally compares against a feature probe fitted directly to its own target, so every encoder appears worse than its own inputs and the diagnostic is uninformative. Probing for slack_v against a matched baseline repairs it. We extract frozen node embeddings from one best-validation-F1 checkpoint per condition, split the evaluation circuits in half, fit a ridge probe from embeddings to slack_v on one half, and score $-\widehat{\mathrm{slack}}$ as a criticality ranker on the other, with the feature baselines of Appendix C.2 refitted on the identical circuits and split.

On the depth-shift set the diagnostic separates conditions exactly at the downstream performance gap. Against a best feature baseline of 0.543 macro AUPRC, the 0.545 of Appendix C.2 refitted on this appendix's half-split, and with forward depth alone at 0.425, three conditions fall below the baseline: the Perceiver at 0.414, the default Node-MLP at 0.491, and the graph-length broadcast at 0.509. These are precisely the three conditions whose downstream macro AUPRC is below 0.60 in Table 4, while the eight clearing the baseline, from 0.577 to 0.738, all have downstream AUPRC of at least 0.827. The classification has no exceptions and the two orderings correlate at Spearman $\rho = 0.82$. Three properties make this non-trivial. The same architectures at random initialization reach only 0.354–0.424, below forward depth alone, so clearing the baseline requires training rather than architecture. The three Node-MLP information conditions rise through 0.491, 0.509 and 0.605, so the intervention of Section 5.1 changes what the frozen representation linearly encodes rather than only the head's operating point. And the Perceiver falls below a single supplied feature, which sharpens Section 5.7.

The classification and the ordering are both stable. Sweeping the ridge penalty over $\alpha \in \{1, 10, 10^2, 10^3, 10^4\}$ leaves the ordering of conditions unchanged and every condition on the same side of the baseline, though absolute transfer rises monotonically with α, from 0.736 to 0.920 for the graph transformer. Reducing the embeddings to six principal components, matching the supplied feature dimension, also preserves the classification and the ordering while raising transfer further, to 0.960. The 128-dimensional probe is therefore under-regularized rather than mis-ranked; consistently, the six-component probe is insensitive to α over this range, since with six standardized features and roughly $46{,}000$ fit nodes even $\alpha = 10^4$ is small relative to the Gram spectrum. This also answers the dimensionality objection, as the advantage over the supplied features survives matching their dimension.

Two scope conditions. The probe is uninformative on the standard set, where forward depth alone already attains 0.833 and the full graph transformer does not clear that baseline despite downstream AUPRC 0.998; with median $L^\star = 8$ there is little descendant-side signal to detect. And the diagnostic detects presence rather than fidelity. GAT-L4 and bidirectional DAGNN are tied on the probe at every α, within 0.006, while their downstream macro AUPRC differs by 0.148, so what distinguishes the DAG-native encoder in Section 5.5 is not captured by a linear readout of slack and is not explained by this measurement.

C.5 Aggregate precision and recall

Table 7 decomposes the three information conditions into precision and recall. The default model operates at high precision and low recall on the standard set and inverts this pattern under depth shift, whereas both intervention conditions are close to balanced on both sets.

Table 7: Aggregate Node-MLP information-condition results. Values are QFT-excluded macro-averages over circuits and population standard deviations over five seeds.

Condition	Set	CP-F1	Precision	Recall	Params
Default summaries	Standard	$.573 \pm .014$	$.819 \pm .015$	$.526 \pm .012$	3,007,563
	Depth shift	$.518 \pm .001$	$.466 \pm .005$	$.633 \pm .008$	3,007,563
Graph-length broadcast	Standard	$.826 \pm .008$	$.850 \pm .008$	$.851 \pm .014$	3,007,948
	Depth shift	$.581 \pm .002$	$.537 \pm .008$	$.673 \pm .011$	3,007,948
Graph-length broadcast + exact reverse depth	Standard	$.974 \pm .008$	$.981 \pm .004$	$.975 \pm .014$	3,007,948
	Depth shift	$.846 \pm .007$	$.829 \pm .025$	$.879 \pm .035$	3,007,948

C.6 Performance across depth-shift path-length bins

Table 8 shows that reverse depth is highly valuable across much of the depth-shift distribution. The benefit is not monotonic in path length: in the extreme upper tail, the learned MLP does not exploit even sufficient inputs as reliably. The likely mechanism is numerical resolution rather than missing information or training coverage. Reverse depth is supplied scaled by $\max(1, L^\star - 1)$, so the model receives $r_v/(L^\star - 1) \in [0, 1]$ with granularity $1/(L^\star - 1)$, and distinguishing $\mathrm{slack}_v = 0$ from $\mathrm{slack}_v = 1$ requires resolving a gap of that size, which is 0.0019 at $L^\star = 523$, through layer normalization and BF16 arithmetic. Two independent signatures are relative rather than absolute: bin-wise F1 declines monotonically in $L^\star$, and the perturbation crossings of Table 6 occur at a fixed fraction of the median longest path on both sets. Training coverage does not track performance here, since the best-performing bin has 16.3% coverage while the 151–250 bin at 4.96% outperforms the 101–150 bin at 4.85%. Supplying the equality residual, or a logarithmic rather than a relative scale, is untested, and it is the run that discriminates the two accounts: if the $251+$ bin recovers under an encoding whose resolution does not degrade with $L^\star$, the deficit is a normalization choice, and the comparator half of the margin in Section 5.5 shrinks accordingly; if it does not, the deficit is a genuine failure of a node-wise predictor to exploit provably sufficient inputs in the deep tail, and the DAG-native margin is attributable to the encoder. We report the mechanism as a hypothesis and the ablation as the test.

Table 9 reports the same condition with precision, recall, and training coverage per bin. Across QFT-excluded depth-shift circuits, mean five-seed Node-MLP F1 is negatively associated with longest-path length (Spearman $\rho = -0.704$) and log gate count ($\rho = -0.648$), while it is positively associated with critical-node fraction ($\rho = 0.350$). A descriptive linear analysis including log path length, log gate count, and critical-node fraction assigns the largest-magnitude standardized coefficient to log path length. These associations do not isolate a causal effect because circuit properties remain correlated.

Table 8: Node-MLP information conditions across QFT-excluded depth-shift path-length bins. Values are mean CP-F1 over five seeds. "Default" is the default summaries condition, "Graph" adds the graph-length broadcast, and "Reverse" additionally adds exact reverse depth. Δ_1 is Graph minus Default and Δ_2 is Reverse minus Graph; both are computed from unrounded values and may differ from the rounded entries by 0.001.

$L^\star$ bin	Circuits	Default	Graph	Reverse	Δ_1	Δ_2
51–75	151	.552	.627	.928	+.075	+.301
76–100	154	.498	.557	.892	+.058	+.335
101–150	166	.483	.542	.813	+.060	+.271
151–250	97	.521	.586	.737	+.066	+.151
251+	28	.646	.670	.713	+.023	+.044

Table 9: Node-MLP performance in the graph-length broadcast plus exact reverse-depth condition, across QFT-excluded depth-shift path-length bins. CP-F1 is reported as a mean and population standard deviation over five seeds; precision and recall are five-seed means. Training coverage is reported for context.

$L^\star$ bin	Circuits	CP-F1	Precision	Recall	Train circuits	Train share
51–75	151	$.928 \pm .014$	.891	.977	32,629	16.34%
76–100	154	$.892 \pm .007$	.873	.920	18,647	9.34%
101–150	166	$.813 \pm .018$	.807	.838	9,681	4.85%
151–250	97	$.737 \pm .021$	.752	.749	9,911	4.96%
251+	28	$.713 \pm .017$	.660	.822	1,656	0.83%

D Local Information-Pathway Ablations

D.1 Reverse-edge communication

Table 10 reports paired changes from removing reverse-edge augmentation while retaining all structural node summaries.

Table 10: Effect of training without reverse-edge augmentation. Values are paired changes relative to full models over all seeds. Structural summaries remain enabled.

Set	Model	ΔCP-F1	ΔPrecision	ΔRecall
Standard	GAT-L4	$-.123 \pm .020$	$-.013 \pm .042$	$-.172 \pm .014$
	GAT-L6	$-.120 \pm .008$	$+.019 \pm .007$	$-.162 \pm .009$
	GIN-L8	$-.130 \pm .012$	$-.068 \pm .064$	$-.154 \pm .011$
Depth shift	GAT-L4	$-.166 \pm .007$	$-.165 \pm .010$	$-.141 \pm .007$
	GAT-L6	$-.190 \pm .011$	$-.188 \pm .019$	$-.170 \pm .004$
	GIN-L8	$-.150 \pm .008$	$-.153 \pm .016$	$-.139 \pm .030$

D.2 Training without grouped structural summaries

Table 11 reports the effect of jointly removing forward depth, in-degree, out-degree, bucketed descendant count, and sequence position from the node inputs while retaining reverse-edge augmentation. The intervention is grouped rather than per-feature, so it establishes the importance of the complete structural pathway without attributing the effect to any single summary. On the standard set the decline is driven primarily by precision for the two GAT configurations, whereas under depth shift precision and recall fall together across all three encoders.

Table 11: Effect of training without grouped explicit structural node summaries. The condition jointly removes depth, degree, descendant-count, and sequence-position summaries. Values are paired changes relative to full models over all seeds.

Set	Model	ΔCP-F1	ΔPrecision	ΔRecall
Standard	GAT-L4	$-.224 \pm .004$	$-.343 \pm .032$	$-.082 \pm .033$
	GAT-L6	$-.178 \pm .050$	$-.231 \pm .077$	$-.086 \pm .051$
	GIN-L8	$-.096 \pm .035$	$-.057 \pm .025$	$-.088 \pm .038$
Depth shift	GAT-L4	$-.202 \pm .003$	$-.172 \pm .005$	$-.171 \pm .008$
	GAT-L6	$-.209 \pm .013$	$-.191 \pm .012$	$-.172 \pm .021$
	GIN-L8	$-.150 \pm .015$	$-.122 \pm .017$	$-.157 \pm .046$

E Depth-Shift Controls

E.1 Full layer sweep

Table 12 reports the complete receptive-field sweep. Depth-shift F1 increases monotonically with layer count for all three encoders, while standard-set F1 is flat or slightly non-monotonic beyond four layers. Because eight propagation steps remain far below the 50 to 522 steps required for exactness, these gains reflect incremental local reach rather than the sufficient-depth regime of Proposition 1.

Table 12: Critical-path F1 versus local message-passing depth. Values are per-circuit means and standard deviations. All entries use five seeds. All aggregates exclude QFT. Layer counts are not parameter-matched: GAT spans 1.83M at two layers to 5.40M at eight, GIN 2.30M to 5.07M, and GCN 2.38M to 4.38M.

Architecture	Layers	Standard	Depth shift
GAT	2	$.789 \pm .018$	$.709 \pm .005$
	4	$.814 \pm .020$	$.790 \pm .004$
	6	$.813 \pm .015$	$.820 \pm .015$
	8	$.802 \pm .005$	$.836 \pm .005$
GIN	2	$.787 \pm .001$	$.723 \pm .005$
	4	$.809 \pm .005$	$.783 \pm .001$
	8	$.810 \pm .010$	$.814 \pm .008$
GCN	2	$.738 \pm .014$	$.703 \pm .002$
	4	$.772 \pm .007$	$.741 \pm .016$
	8	$.755 \pm .004$	$.790 \pm .010$

E.2 Fine path-length bins

Table 13 resolves the two-to-eight-layer gain within path-length bins. The gain is positive in every bin for all three encoders and peaks in the 101–150 band rather than in the extreme upper tail.

Table 13: Depth-shift F1 gain from two to eight local layers within path-length bins.

Model	51–75	76–100	101–150	151–250	251+
GAT	+.118	+.133	+.148	+.111	+.085
GIN	+.076	+.099	+.105	+.084	+.057
GCN	+.074	+.082	+.103	+.092	+.066

Maximum path length varies across qubit-count bands, but positive layer trends remain within populated bands. The results therefore support a local-reach explanation beyond a simple qubit-count confound.

F DAG-NATIVE PROPAGATION

Configuration and faithfulness. Two DAGNN layers at hidden width 276, additive query/key attention with GRU combination, bidirectional. Width was selected by search rather than tuned: at width 192 the implementation has 1,713,295 parameters, -43.2% against the GAT-L4 reference, and 276 is the only even width inside a $\pm 2\%$ tolerance, giving 3,024,602 parameters at $+0.23\%$. Even width is required by the inherited sinusoidal positional encoding. Each DAGNN layer is a full-diameter topological sweep, so the layer count is refinement depth rather than receptive-field radius; we verify by perturbation that a single layer propagates to all 16 positions in both directions, with an edge-free control giving exactly 0.000 change at every unperturbed node. The broadcast projection is bias-free $\mathrm{Linear}(2 \to H)$ plus a scalar gain, adding exactly $2H + 1 = 553$ parameters, and its second channel is held at zero so projection shape matches the graph-length-only condition. The gradient on that zero channel is exactly 0.0, and across all five seeds the $L^\star$ channel norm grows from 1.37 to 7.67–8.09 while the zero-input channel decays to 0.85–0.87 under weight decay alone. The same diagnostic on the GAT-L4 broadcast condition of Section 5.3 gives a Xavier-uniform 1.99 growing to about 4.9, a gain rising from 1.00 to about 1.29, and the zero-input channel decaying to 0.92–0.97, on every one of five seeds.

Cone coverage. Node v's representation is a function of $\{v\} \cup \mathrm{anc}(v) \cup \mathrm{desc}(v)$, whereas $L^\star$ is a maximum over all nodes including parallel qubit branches that never interact with v. Writing $\text{cone-max}(v) = \max_{u \in \mathrm{cone}(v)} s_u$, the deficit $L^\star - \text{cone-max}(v)$ is computable in $O(|V| + |E|)$ by two running maxima and is zero exactly when the cone contains a critical node. Because critical nodes are trivially in their own cone, we report the statistic over non-critical nodes, of which there are 3,077 on the standard set and 255,295 under shift. Two bounds are relevant. Taking the cone as $\mathrm{anc}(v) \cup \mathrm{desc}(v)$ gives a positive deficit at 37.4% of non-critical standard-set nodes against 6.1% under shift. The forward and backward sweeps of the evaluated encoder do not exchange information before the final concatenation, however, so the ancestor-side running maximum is not in fact reachable and the operative quantity is the descendant-side maximum alone; this raises the deficit to 45.0% and 19.2%. Mean cone size is 0.525 of the graph on the standard set against 0.675 under shift, so reconvergence in deeper circuits is the mechanism. Under the descendant-side bound the deficit is 31.4%, 23.2%, 14.3%, 18.5% and 3.8% across the five depth-shift $L^\star$ bins of Table 8, so it is not monotone in path length and $L^\star$ remains unreachable for a minority of nodes even in the deepest circuits. The broadcast is therefore an availability test on both sets, more strongly on the standard set, and additionally a realizability test wherever $L^\star$ is present in the cone, since softmax-weighted sum must approximate the max that Proposition 1 requires. The same pass verifies, on all 103 standard and 596 depth-shift circuits, zero violations of Equation 2, zero disagreements between the threshold rule and the stored labels, zero deviations of stored depth from the exact longest-path layering $d_v = 1 + \max_{p \in \mathrm{pa}(v)} d_p$, and zero negative deficits.

Batching control. Activation memory scales as batch size times $L^\star$, so the DAG-native arm uses length-bucketed batching with batch size 64, 32 and 16 for $L^\star$ in $[4, 127]$, $[128, 315]$ and $[320, 546]$, affecting 8.4% of the corpus. Bucketing also makes the batch-maximum normalizer of the depth-regression target a near-exact per-circuit proxy for $L^\star$, which under random shuffling it is not (Appendix I). Two controls bound this. First, a behavioural test of whether the channel is used at all. Because the depth target is d_v over a batch maximum, a model trained under shuffling learns an approximately *absolute* depth scale, whereas one that internalizes a per-circuit $L^\star$ learns a *relative* one. Within a single circuit the two are indistinguishable, both proportional to d_v, but the per-circuit gain separates them: fitting $\widehat{d}_v \approx a_c d_v$ within each circuit and regressing $\log a_c$ on $\log(L^\star - 1)$ across circuits gives slope 0 for a constant normalizer and -1 for a per-circuit one. On the depth-shift set GAT-L4 trained under shuffling gives slope -0.17 at $R^2 = 0.15$, whereas the same architecture trained under the bucketed sampler gives -1.12 at $R^2 = 0.97$, with the correlation between its depth output and relative depth $d_v/(L^\star - 1)$ rising from 0.766 to 0.923. Bucketing therefore does install a per-circuit $L^\star$ representation: the channel is real and it is exploited. Second, whether that matters for the predicate. It does not help here. Retraining GAT-L4 with the identical sampler gives paired F1 differences of -0.028 $[-0.053, -0.007]$ on the standard set and -0.015 $[-0.017, -0.013]$ under shift, negative on every seed, with standard-set macro AUPRC unchanged at 0.978 and depth-shift macro AUPRC falling from 0.827 to 0.809. This is what Section 5.2 predicts, since $L^\star$ is a per-circuit constant and cannot reorder nodes within a circuit, so an implicitly supplied one can move

the operating point but not the ranking. The measured effect is net rather than decomposed, since reduced within-batch gradient diversity and per-batch class imbalance act in the same direction. Applying the same diagnostic to the DAG-native arm is less favorable to it. Bidirectional DAGNN gives slope -0.99 at $R^2 = 0.99$ with relative-depth correlation 0.986, so it too holds a per-circuit $L^\star$, obtained without any broadcast. This is consistent with the cone analysis above, which finds $L^\star$ present in the descendant-side cone at 80.8% of non-critical depth-shift nodes, and it means that "no $L^\star$" must be read throughout as "no explicit $L^\star$ broadcast" rather than as no access by any route. The three Node-MLP conditions give slopes between -0.22 and -0.23 at $R^2 \leq 0.44$ and so serve as a negative control, and the full graph transformer gives -0.44 at $R^2 = 0.71$ without a broadcast, which we return to in Appendix H. The decisive experiment remains to retrain bidirectional DAGNN with the depth-regression normalizer fixed to a global constant, or with that auxiliary task ablated, under the identical sampler; we have not run it. What these controls establish is that the implicit channel exists, that it is acquired by models able to acquire it, and that in the one case where its effect on this predicate can be measured directly it lowers rather than raises node ranking.

Table 14: Standard-set macro F1 within circuit-depth bins, five seeds as in Table 1. The DAG-native deficit is confined to the 38 circuits with $L^\star \leq 5$, which macro-averaging weights equally with the largest circuits; this also accounts for the gap between macro and pooled F1 for DAGNN in Table 4.

Condition	$L^\star \leq 5$ (n=38)	6–10 (n=24)	11–25 (n=23)	≥ 26 (n=18)
DAGNN, bidir + $L^\star$	.685	1.000	.999	**.994**
DAGNN, bidir	.478	.952	.985	.993
GAT-L4	.563	.973	.997	.897
GAT-L8/H138, matched	.577	.969	.993	.898

G Perceiver Controls

Table 15 reports the latent-array sweep. Changing M by $16\times$ changes depth-shift F1 by only about 0.02. Relative-depth clipping to $[-64, 64]$ is unlikely to explain this result: median absolute edge depth difference is one, the 99th percentile is 17, and only 0.05% of edges exceed the clipping range.

Table 15: Perceiver critical-path F1 versus latent-array size. Entries without uncertainty are single runs.

Configuration	Standard	Depth shift
Perceiver, $M = 32$	.788	.390
Perceiver, $M = 64$	.732	.388
Perceiver, $M = 128$	.726	.393
Perceiver, $M = 256$	$.748 \pm .010$	$.408 \pm .011$
Perceiver, $M = 512$	.730	.403

The single-seed 8.4M-parameter Perceiver control reaches approximately 0.745 QFT-excluded standard F1. This control tests one larger configuration and does not establish a universal scaling conclusion.

H Graph-Transformer Conditions

Why supplied structure is worth so much here. All-pairs attention with a relative-depth bias is unusually well matched to Equation 1. A single attention layer can evaluate $\max_u d_u = L^\star - 1$, since every node attends to every other and the bias resolves depth differences pairwise, and the same mechanism supports descendant-side extrema through the directed edge-type term. The full graph transformer therefore possesses both quantities that the Node-MLP interventions of Section 5.1 show to be necessary, which accounts for near-ceiling performance in the full condition and for the size of the drop when the relative-depth bias is removed. The depth-head diagnostic of Appendix F supports this directly: on the depth-shift set the full graph transformer attains a per-circuit depth-gain slope of

Table 16: Graph-transformer results over five seeds. "Full" uses structural node summaries and structural attention biases. "Reduced-input" is trained without node summaries or relative-depth bias but retains semantic inputs, wire identity, and directed edge-type bias. "Inference masking" removes the complete structural pathway only at inference.

Set	Condition	CP-F1	Precision	Recall
Standard	Full	$\mathbf{.981 \pm .008}$	$.974 \pm .013$	$.993 \pm .003$
	Reduced-input	$.790 \pm .044$	$.751 \pm .060$	$.904 \pm .034$
	Inference masking	$.391 \pm .177$	$.341 \pm .120$	$.555 \pm .280$
Depth shift	Full	$\mathbf{.982 \pm .002}$	$.980 \pm .004$	$.986 \pm .002$
	Reduced-input	$.720 \pm .006$	$.692 \pm .005$	$.773 \pm .014$
	Inference masking	$.392 \pm .055$	$.302 \pm .046$	$.769 \pm .074$

-0.44 at $R^2 = 0.71$, the largest magnitude of any condition that receives neither an $L^\star$ broadcast nor the bucketed sampler, against -0.17 for GAT-L4 and between -0.22 and -0.23 for the three Node-MLP conditions. It partially internalizes a per-circuit graph-length scale without being given one. The reduced-input condition cannot be read on this diagnostic, since with structural summaries removed its depth output barely tracks depth at all, correlating 0.044 with d_v on the standard set, so its fitted slope describes noise rather than normalization. Table 16 reports the three graph-transformer conditions. The reduced-input graph transformer contains $3{,}320{,}289$ trainable parameters. Its standard seed-level F1 ranges from 0.738 to 0.870, while depth-shift F1 ranges from 0.710 to 0.728. The inference-masking condition is bimodal rather than stably degraded. Standard-set seed-level F1 values are 0.203, 0.214, 0.340, 0.594, and 0.605, and depth-shift values are 0.322, 0.343, 0.388, 0.451, and 0.455, so the tabulated means near 0.39 average over two distinct behaviours and should not be read as characteristic operating points. This condition changes the model's input distribution and removes structural attention bias that remains available under reduced-input training. Inference masking and reduced-input training therefore differ in both adaptation and retained information; their difference cannot isolate either factor separately.

I Implementation Details

Descendant summary. The implemented descendant target is the capped recursion $c_{\mathrm{desc}}(v) = \min\big(|V|,\ \sum_{w \in \mathrm{ch}(v)} (1 + c_{\mathrm{desc}}(w))\big)$, with the empty sum zero so that $c_{\mathrm{desc}}(v) = 0$ at every sink. It is chosen because a single reverse-topological pass computes it, whereas $|\,\mathrm{desc}(v)|$ requires transitive closure. On reconvergent DAGs it may count a downstream node through multiple child branches before capping, so it is a descendant-derived summary rather than necessarily the unique-set cardinality. It is not equivalent to reverse longest-path depth: two nodes with identical c_{desc} can have different r_v and different labels, which is why the bucketed count in the default feature set does not make the no-reverse-depth conditions descendant-sufficient.

Objective definitions. Masked gate type and gate family are cross-entropy at masked positions with label smoothing 0.05, and gate-type classes are weighted by inverse square-root frequency, normalized to mean one and clipped to $[0.2, 10]$. Masking replaces the gate-type index at selected positions with a dedicated learned mask embedding, the extra row indexed $\mathrm{num_gate_types}$ in a vocabulary-plus-one table. The gate-family input is *not* masked, so the family task is close to trivial; it accordingly carries weight 0.01, falling to 0.001 once validation family accuracy exceeds 0.97. The masking rate follows a fixed epoch-indexed schedule rising to a maximum of 0.30 during training and is held at 0.15 for every validation pass. Depth regression targets d_v divided by the largest valid depth *in the batch* rather than in the circuit: under the random shuffling used by every condition except the DAG-native arm this normalizer is approximately the expected maximum over 64 draws from the corpus depth distribution, hence effectively constant and not a per-circuit function of $L^\star$. Appendix F measures what changes under length-bucketed batching, where it is not. Descendant-count regression targets $\log(1 + c_{\mathrm{desc}}(v))$, likewise divided by its batch maximum. Critical-path classification uses focal loss with $\gamma = 1.0$ and $\alpha = 0.50$, or $\alpha = \mathrm{clip}(1 - \pi,\ 0.2,\ 0.8)$ for corpus critical-node rate π where the data-driven option is enabled. Next-gate prediction is cross-entropy over gate types at the next *sequence position* in the stored gate ordering, not over graph children, so it

is a sequence-model auxiliary rather than a structural one. Sinusoidal sequence position encodes the index of the gate in that same stored ordering, added through a learned scale initialized to 0.3. Task weights were fixed once during preliminary development rather than tuned per condition, and every architecture and information condition reported here is trained under identical objective settings.

Pairwise structural bias in the global encoders. The graph transformer computes all-pairs attention over N nodes and adds directed edge-type and clipped relative-depth biases directly to the $N \times N$ logit matrix, so each pair (u, v) receives its own bias term. The Perceiver attends between M latents and N nodes, so an $N \times N$ bias has no direct analogue. Our implementation computes the same $N \times N$ edge-type and relative-depth bias tensor and then reduces it along the query axis to one scalar per key per head, $\bar{b}_j = \frac{1}{N} \sum_i b_{ij}$, which is added to every latent's attention logit over node j. Pairwise structure therefore reaches the Perceiver only as a per-node scalar salience, identical for all latents, whereas the graph transformer retains the full pair-resolved term. The two global encoders are consequently not matched on pairwise structural information, and the gap in Section 5.7 cannot separate a latent-routing limitation from this weaker bias channel; the latent sweep varies capacity rather than the channel, so it does not adjudicate between them either.

Scope conditions on the DAG-native arm. The DAG-native conditions use five seeds at a single layer count and width, mean pooling rather than the original's max over target nodes, and softmax-weighted-sum rather than the max aggregation that Proposition 1 requires. The arm also varies full-diameter propagation, relation-specific parent and child channels, and GRU combination together. A parameter-matched local encoder with relation-specific channels but finite reach would separate the first factor from the second and is not run here, so Section 5.5 supports a sufficiency claim and not an attribution.

J Routing-Overhead Triage

J.1 Protocol

We generate 800 circuits with 16–28 qubits, depths from 20 to 80, and up to $1{,}800$ gates. Each circuit is transpiled using Qiskit's SABRE layout and routing (Li et al., 2019; Javadi-Abhari et al., 2024) against line, ring, grid, heavy-hex-like, and sparse-random coupling maps.

Targets are

$$\text{depth inflation} = \frac{d_{\text{post}}}{d_{\text{pre}}}, \qquad \text{two-qubit inflation} = \frac{n_{\text{2Q,post}}}{n_{\text{2Q,pre}}}.$$

We compare:

1. a 15-dimensional handcrafted circuit-and-topology descriptor;
2. pooled pretrained GAT-L6 embeddings plus topology descriptors;
3. architecture-matched random-initialized GAT embeddings plus topology descriptors.

The pretrained feature concatenates mean-, max-, and standard-deviation-pooled 128-dimensional node embeddings, producing 384 circuit features, followed by six topology features. A standardized ridge probe is trained on 600 circuits and evaluated on 200, with five encoder seeds and a fixed circuit split. These routing circuits are generated independently of the pretraining corpus. Two caveats bound the comparison. The embedding probes receive 390 features against 15 for the handcrafted descriptor, and because the architecture-matched random-initialized embeddings also exceed the handcrafted baseline, part of that margin is probe capacity rather than representation quality; the pretrained-versus-random-initialization contrast is dimension-matched and is therefore the controlled comparison, while the handcrafted comparison should be read as a reference point only. The handcrafted descriptor and the six topology features are enumerated in the accompanying configuration files.

J.2 Results

Pretrained embeddings obtain per-topology Spearman correlations from 0.74 to 0.93. They improve over random initialization by 0.06–0.15 and over handcrafted features by 0.06–0.19 in every

topology–target cell. For two-qubit inflation, pretrained embeddings obtain $R^2 = 0.79 \pm 0.004$, compared with 0.71 ± 0.015 for random initialization and 0.64 for handcrafted features. For depth inflation, the corresponding values are 0.89 ± 0.008, 0.79 ± 0.018, and 0.84.

Removing topology descriptors preserves within-topology ranking but prevents reliable cross-topology selection. Thus, embeddings provide circuit information, while topology descriptors are required to compare candidate devices.