STRUCTURE-ALIGNED SYNTHETIC PRIORS FOR TABULAR FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular foundation models such as TabPFN, TabICL, and MITRA are pre-trained on millions of synthetic datasets sampled from randomly generated Structural Causal Models (SCMs). In the dominant paradigm the SCM is realised as a randomly wired multi-layer perceptron (MLP) whose activations are read off as features and targets, so the dependency graph of the training data is an artifact of MLP connectivity rather than a deliberate choice. We argue that this *topology* is an independent design variable, and that it diverges systematically from real causal structure. Comparing $\sim$36,000 TabICL SCMs against 33 real-world causal graphs of three independent provenances, we find that MLP priors produce conditional dependence structures an order of magnitude wider: median Markov-blanket width 43.8 versus 4.0, maximum parent count 31 versus 3, maximum minimum-d-separator cardinality 20 versus 3. This is not a restatement of graph size—MLP blanket width *scales* with node count ($\propto N^{0.73}$) whereas real-world width is size-invariant ($N^{0.06}$, n.s.), and the gap persists at $10.2\times$ within a matched size band. We therefore introduce BN-ICL, a prior whose dependency graph is an explicit Bayesian network matched to real structural statistics, and train in-context learners *from scratch* under an ablation in which the *only* difference is DAG topology. On an augmented suite of 219 real classification datasets (TALENT plus a high-dimensional OpenML expansion specified in advance), the BN prior improves mean accuracy and log-loss, but not uniformly: the effect is concentrated in, and grows with, feature dimensionality, and the priors are indistinguishable on the median dataset. Above 100 features the BN prior wins by $+1.5$–1.6 percentage points (bootstrap 95% CI excluding zero, Holm-corrected $p < 0.05$), replicating across two independently trained prior configurations and on external datasets disjoint from TALENT—ruling out a benchmark-composition artifact—and surviving truncation of those tables to the training feature width when the retained features are informative, which rules out width extrapolation as the explanation. Low- and mid-dimensional gains are smaller or fragile. DAG topology is thus a meaningful, under-explored axis of prior design, sharpest where the structural discrepancy is largest.

1 INTRODUCTION

Foundation models have transformed natural language and vision, but the tabular domain—the dominant data type in banking, e-commerce, healthcare, and manufacturing—has only recently acquired comparable pre-trained models. TabPFN Hollmann et al. (2023; 2025), TabICL Qu et al. (2025), and MITRA Zhang et al. (2025) share a common recipe: because real tabular datasets are scarce and heterogeneous, they pre-train a transformer via in-context learning (ICL) on *millions of synthetic datasets* sampled from randomly generated Structural Causal Models (SCMs). The premise is that data everywhere arises from causal mechanisms, so a sufficiently diverse family of random SCMs should span the distribution of real-world problems, and a transformer trained to Bayesian-infer under that prior will generalise to real tables at inference time without weight updates.

The performance of such a Prior-Fitted Network is therefore determined almost entirely by *how well the synthetic prior matches real tabular data*. This places the design of the SCM generator at the centre of the enterprise. In the dominant paradigm, the SCM is instantiated as a randomly initialised MLP: root "causes" are pushed through several dense (block-sparsified) layers, and a random subset

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.

of intermediate activations are designated as features X and targets y. The dependency graph of the resulting data is thus an *artifact of MLP connectivity*, not a deliberately chosen causal topology.

We make three contributions. **(i)** We quantify the structural gap between MLP-induced SCM graphs and real causal graphs. Comparing $\sim$36,000 TabICL SCMs against a corpus of 33 real-world causal graphs of three independent provenances, we find that MLP priors produce graphs with an order-of-magnitude wider conditional dependence structure—larger Markov blankets, higher parent counts, and larger minimum d-separating sets—together with weaker hub concentration. We further show that this discrepancy is a property of *topology* rather than of graph size, since blanket width grows with node count under the MLP prior but is size-invariant in real graphs (Section 5). **(ii)** We introduce BN-ICL, a prior whose dependency graph is an explicit Bayesian network sampled to match the structural statistics of real causal graphs, and we design a *controlled ablation* in which BN-ICL and a faithful MLP prior differ in *nothing but DAG topology*—identical feature mechanisms, labelling, normalisation, architecture, optimiser, data budget, and seed (Section 6). **(iii)** Training both from scratch and evaluating on an augmented suite of 219 real classification datasets, we show that DAG topology alone produces a statistically significant improvement that is *concentrated in, and grows with, feature dimensionality*: on high-dimensional tables the BN prior wins significantly and reproducibly, an effect that replicates across two prior configurations and across independent data sources (Section 7).

We are explicit about scope. Our models are trained at deliberately reduced scale to isolate the topological variable under a fixed compute budget, so both arms trail the published TabICL checkpoint in absolute terms; our claim is not a new state of the art but a *guideline* that the topology of the synthetic prior is a consequential, tunable design choice, most visibly in the high-dimensional regime.

2 Related Work

Tabular foundation models. TabPFN Hollmann et al. (2023) is the first transformer-based tabular foundation model trained by ICL on purely synthetic SCM data; TabPFN v2 Hollmann et al. (2025) adds a deeper architecture, larger dataset support, and two-dimensional attention. TabICL Qu et al. (2025) introduces a column-then-row attention factorisation that scales ICL to large tables. MITRA Zhang et al. (2025) improves the synthetic prior by mixing generator families, and TabDPT Ma et al. (2025) instead combines ICL with self-supervised learning on real data. Concurrent work extends both lines at substantially larger scale than we can match here—TabICLv2 Qu et al. (2026) and TabPFN-3 Grinsztajn et al. (2026). Our MLP arm reproduces the generator of TabICL Qu et al. (2025); the SCM-based models in this family share the MLP-induced-graph construction, and our analysis concerns that construction rather than any particular checkpoint.

Priors as the object of study. Because a Prior-Fitted Network learns to perform posterior inference under its training prior, the prior *is* the model's inductive bias, and several recent results turn on changing it: MITRA's gains from richer generator mixtures corroborate that prior design drives performance. TabForest and TabForestPFN den Breejen et al. (2024) are an instructive contrast, fitting an overfitted decision tree to Gaussian noise to obtain datasets that are deliberately *unrealistic* but have complex decision boundaries. That generator has no dependency graph at all, so it is orthogonal to our axis rather than an alternative point on it—and notably, TabForestPFN's strongest configuration *mixes* the forest generator with the MLP-SCM generator, so the structural properties we document apply to half of its pre-training distribution. Where their axis is decision-boundary complexity, ours is dependency topology: we hold the generative *mechanism* fixed and vary only the *graph*, measuring the marginal contribution of structure.

3 Background on Causal Graphs

We use standard causal-graph terminology Pearl (2009); full definitions, the three path configurations (chain, fork, collider), and d-separation are in Appendix A. A causal graph $G = (V, E)$ is a DAG whose nodes are random variables and whose edges denote direct causal influence, with the

joint factorising via the Causal Markov Condition,

$$P(X_1, \ldots, X_N) \;=\; \prod_{i=1}^{N} P(X_i \mid \mathrm{Pa}(X_i))\,. \tag{1}$$

Two derived quantities carry our analysis. The *Markov blanket* $\mathrm{MB}(v) = \mathrm{Pa}(v) \cup \mathrm{Ch}(v) \cup \{\mathrm{Pa}(c) : c \in \mathrm{Ch}(v)\}$ is the minimal set rendering X_v independent of all other variables, so $P(X_v \mid \mathrm{MB}(v))$ has dimension $|\mathrm{MB}(v)|$; and the *minimum d-separator* $\mathbf{Z}_{\min}(u, v)$ is the smallest set rendering a non-adjacent pair conditionally independent. Both measure how far a joint distribution decomposes into low-dimensional pieces, and hence how sample-efficiently it can be learned: the factorisation (1) is sample-efficient precisely when parent sets are small, since for discrete variables with s states each conditional carries $O(s^{|\mathrm{MB}(v)|})$ free parameters.

4 The Topology of MLP-Induced Priors

An SCM realised as an MLP begins from a layer-wise fully connected topology and stochastically sparsifies it via block-wise dropout or edge pruning. Two structural properties survive this process. First, edges only ever connect adjacent layers, so the graph is a strict feed-forward cascade: a node in layer t can influence layer $t+n$ only through the intervening layers. Assuming, as in the base generator, that there are no skip or residual connections, this imposes a rigid Markov chain across layers,

$$L_{t+n} \perp\!\!\!\perp L_{t-n} \mid L_t, \qquad \forall\, n \geq 1, \tag{2}$$

so any long-range dependency in the true generating process must be routed through, and diluted by, every intervening layer. Second, even after aggressive pruning, hidden layers are initialised wide, so the surviving parent set of a node, $|\mathrm{Pa}(h_j^{(t)})| = K_j < N_{t-1}$, remains large relative to a sparse real-world graph, and the local conditionals $P(h_j^{(t)} \mid \mathrm{Pa}(h_j^{(t)}))$ retain a high-dimensional, block-structured dependence on the preceding layer.

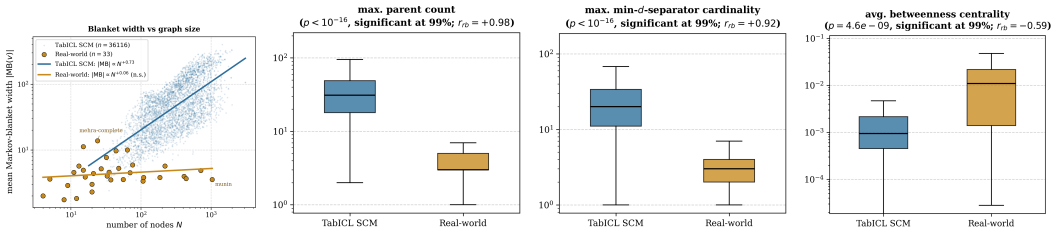
We note for precision that this is a claim about parents *per node*: because normalised density $2|E|/(N(N-1))$ carries a factor $1/N$, the much larger MLP graphs can simultaneously be wider per node and *lower* in normalised density, and in our measurements they are (Section 5). It is the per-node quantity that sets the dimension of each conditional, and hence learnability. Real causal systems are overwhelmingly sparse and bounded-parent; MLP priors generate data whose conditionals are wide and whose conditional independences, though numerous, require conditioning sets that grow with layer width to certify (Section 5.2). This is a purely *topological* argument: in Section 6 we hold the mechanisms fixed and vary only the topology, so any downstream difference is attributable to structure alone.

5 Structural Discrepancy Between SCM and Real-World Graphs

To quantify the gap, we sampled 36,116 SCMs from the TabICL generator ($\mathcal{G}_{\mathrm{SCM}}$) and compared them against a corpus of 33 real-world causal graphs ($\mathcal{G}_{\mathrm{RW}}$) of three independent provenances: (i) 16 expert-elicited benchmark Bayesian networks from the `bnlearn` repository Scutari (2010); (ii) 5 reference networks learned from real data by their original authors; and (iii) 12 networks whose structure we learned ourselves from public tabular datasets using score-based structure learning. Full provenance, the learning procedure, and per-graph node and edge counts are in Appendix C. The three provenances matter: expert networks are not subject to the sparsity bias of a structure-learning algorithm, while learned networks are not subject to the parsimony preferences of a human draughtsman, and *all three exhibit the same profile*. Unlike the SCM generator, which can emit unlimited synthetic graphs, validated real-world causal graphs are a scarce resource; our corpus is therefore modest in size but deliberately spans 4 to 1041 nodes and a wide range of application domains, and it includes dense counter-examples.

For each graph we computed size (node and edge counts), parent counts (maximum and mean), Markov-blanket width, the cardinality of minimum d-separating sets, and degree, closeness, and betweenness centrality. All metric definitions and normalisations are in Appendix A.4. Both load-bearing quantities (Section 3) are *per-node* rather than counts, which is what makes them comparable

across a corpus spanning two orders of magnitude in size. The four most informative comparisons appear in Figure 1; all remaining metrics, including every non-significant one, are in Appendix A.5.



(a) (b) (c) (d)

Figure 1: TabICL SCMs ($n = 36{,}116$, blue) vs. real-world causal graphs ($n = 33$, orange). **(a)** Markov-blanket width against node count: MLP width *scales* with graph size ($\propto N^{0.73}$) while real-world width is size-invariant ($N^{0.06}$, n.s.), so the gap is topological, not a size effect. **(b–d)** Real-world graphs have $10.3\times$ lower maximum parent counts, $6.7\times$ smaller minimum d-separators, and higher betweenness centrality (hub concentration). Mann–Whitney p and rank-biserial r_{rb} are annotated per panel. Blanket-width and parents-per-node boxplots, and all non-significant metrics, are in Appendix A.5.

Findings. The comparison yields a coherent picture across five significant axes, all with very large effect sizes (rank-biserial $|r_{rb}|$). **(1) Conditional width.** Real-world graphs have a median Markov-blanket width of 4.0 against 43.8 for TabICL—a $10.9\times$ gap with $r_{rb} = 0.979$, i.e. near-total separation of the two distributions ($p < 10^{-21}$; Appendix A.5). The widest real-world graph in the corpus (MEHRA, $|\mathrm{MB}| = 13.8$) still lies below the *median* TabICL SCM. **(2) Parent counts.** Maximum parent counts differ by $10.3\times$ (median 3 vs. 31) and mean parent counts $|E|/N$ by $5.6\times$ (1.52 vs. 8.50; Appendix A.5), the hallmark of sparse, bounded-parent causal structure. **(3) Conditioning complexity.** The maximum minimum-d-separator cardinality is 3 for real-world graphs against 20 for TabICL ($6.7\times$, $p < 10^{-19}$); the corresponding means are 1.09 and 5.05. Since the statistical cost of a conditional-independence judgement grows exponentially in the size of the conditioning set, this is the decision-relevant contrast—in the binary case, roughly 2^3 against 2^{20} configurations. **(4) Hub concentration.** Real-world graphs show significantly *higher* average betweenness centrality (0.011 vs. 0.001; $r_{rb} = -0.589$, $p < 10^{-8}$) and maximum betweenness ($p < 10^{-11}$), indicating that influence is routed through a small number of hub/bottleneck nodes rather than spread across a dense, homogeneous mesh. **(5) Scaling, not size.** The gap is not a restatement of the fact that real graphs are smaller. Markov-blanket width in the MLP prior *grows* with node count, $|\mathrm{MB}| \propto N^{0.73}$ ($r^2 = 0.45$, $p < 10^{-300}$), whereas in real graphs it is statistically size-invariant, $\propto N^{0.06}$ ($r^2 = 0.03$, $p = 0.37$): MUNIN has 1041 nodes and a mean blanket of 3.5, and PIGS has 441 nodes and 3.7 (Figure 1a). Restricting both corpora to the matched band $N \in [100, 500]$ leaves the gap at $10.2\times$ ($r_{rb} = 0.999$, $p = 2.3 \times 10^{-5}$), and restricting the real-world corpus to $N \geq 20$—which removes every small learned network—leaves it at $10.0\times$ (Appendix A.5).

Two metrics that do not support a claim. We report two negative results explicitly, since both are frequently cited and both would mislead here. First, normalised density $2|E|/(N(N-1))$ and average degree centrality (which are proportional) are *not* significant at 99% ($p = 0.031$) and run *opposite* to the per-node measures (real-world 0.095 vs. TabICL 0.064)—an arithmetic consequence of the $1/N$ factor, since at fixed parents-per-node a larger graph realises a smaller fraction of its possible edges, and the real-world graphs are $9\times$ smaller. We therefore define "denser" throughout as mean parent count $|E|/N$ and rest no claim on the N^2-normalised quantities. Second, the corpora's internal structural diversity is directionally consistent with the MLP corpus being more redundant, but does not survive matching on corpus size and node count, so we do not claim it (Appendix A.6). Closeness centrality is likewise not significant ($p > 0.87$).

5.1 MEASURING CONDITIONING COMPLEXITY

Markov-blanket width is computed exactly on every graph: $\mathrm{MB}(v)$ is precisely the neighbourhood of v in the moral graph of G, so the full distribution follows from a single moralisation pass per

graph (Appendix A.3). For minimum d-separator cardinality we use the characterisation of Tian et al. (1998): for a non-adjacent pair (u, v), a minimum-cardinality d-separator is a minimum vertex cut between u and v in the moral graph of the subgraph induced by the *ancestral set* $\mathrm{An}(\{u, v\})$. The ancestral restriction is essential rather than an optimisation: moralising the entire DAG marries co-parents everywhere, including pairs that the ancestral construction provably cannot connect (Appendix A.3), which would systematically understate separability. We sample non-adjacent pairs per graph—exhaustively for the real-world corpus, and 150 pairs on each of 3,000 uniformly sampled SCMs—and report per-graph summary statistics. Separation-based distances between causal graphs have been studied directly by Wahl & Runge (2025), whose framework motivates treating separation structure, rather than edge overlap, as the object of comparison. No graph is excluded from any analysis and no per-graph time budget is applied.

We deliberately do *not* report the *number* of d-separable pairs in a graph. Every non-adjacent pair of a DAG is d-separable, so that count equals $\binom{N}{2} - |E|$ exactly and its normalised form is one minus the density—a function of size and density carrying no topological information, and in our corpora not significant ($p = 0.031$). Appendix A.3 gives the derivation and the measured values; this is why we formulate the comparison in terms of separator *cardinality* and blanket *width* instead.

5.2 Why MLP priors widen conditionals: a structural account

The empirical gaps in Figure 1 follow from the two generators' topologies, and—crucially—from how they behave as the node count N grows.

Conditioning complexity is bounded by parent count.

Theorem 1 (Separator bound)**.** *Let u, v be non-adjacent in a DAG G. Then* $|\mathbf{Z}_{\min}(u, v)| \leq \max_{w \in V} |\mathrm{Pa}(w)|$.

The proof is in Appendix A.3; it follows from applying the local Markov property to whichever of the two nodes is not a descendant of the other. Theorem 1 links the two most robust measurements in Figure 1: bounded parent sets *guarantee* cheap conditioning. It holds with zero violations across all 3,033 graphs on which we compute both quantities. The bound is not tight—real-world graphs have maximum parent count 3 and maximum separator cardinality 3, while TabICL SCMs have 31 and 20—but it establishes that the parent-count gap and the conditioning-complexity gap are not independent findings: the former causes the latter.

Blanket width: bounded in the BN prior, growing in the MLP prior. With k_{pa} parents, k_{ch} children, and $k_{\max} = \max_w |\mathrm{Pa}(w)|$, the definition of the Markov blanket (Section 3) gives

$$|\mathrm{MB}(v)| \;\leq\; k_{\mathrm{pa}} + k_{\mathrm{ch}} + k_{\mathrm{ch}}\,(k_{\max} - 1) \;=\; O(k_{\mathrm{ch}}\, k_{\max})\,. \tag{3}$$

The two generators sit at opposite ends of this expression. In the BN prior, in-degree is capped at a constant M matched to the real-world profile and the expected out-degree is $\bar{k} = O(1)$ because the edge budget is linear in N; hence $|\mathrm{MB}| = O(1)$, *independent of* N. In a block-sparsified MLP over L layers of width H, a mid-layer node has $\Theta(k)$ parents, $\Theta(k)$ children, and $\Theta(k)$ co-parents through each child, with k growing in the layer width H; since $N = n_{\mathrm{c}} + LH$, blanket width therefore grows with N. This is exactly what we measure: $|\mathrm{MB}| \propto N^{0.73}$ for the MLP prior against $N^{0.06}$ (not significant) for real graphs (Figure 1a).

Equation (11) in Appendix B.4 is the complementary statement for the BN prior: a linear edge budget gives average degree centrality $\mu_{C_D} = 2\bar{k}/N$, decreasing in N. We stress that Eq. (11) and Eq. (3) measure different things—the former a fraction of possible edges, the latter an absolute conditional dimension—and that only the latter governs sample complexity. This is why the two normalisations disagree in sign in our data (Section 5).

The layer constraint, restated as a cost. The strict feed-forward wiring of an MLP prior forbids edges between non-adjacent layers and so imposes the layer Markov chain of Eq. (2). Every pair of nodes two or more layers apart is therefore d-separable. We emphasise that this is not by itself a discriminating property: every non-adjacent pair in *any* DAG is separable, so what distinguishes the priors is not *whether* a pair can be separated but at what *cost*. Under the layer constraint, separating such a pair requires conditioning on $\Theta(k)$ variables with k growing in H—because moralisation

turns each densely-wired layer into a near-clique—against at most $M = O(1)$ in the BN prior. We measure 20 against 3. Because the emitted feature count grows with H, the MLP prior's conditionals widen precisely as the table widens, while the BN prior's stay bounded. This predicts that the BN prior's advantage should be largest in the high-dimensional regime, which is what we observe in Section 7.

6 Proposed Approach: A Controlled Topology Ablation

Our goal is to test a single causal claim: *does replacing MLP-induced graph topology with real-world-matched Bayesian-network topology improve downstream performance, all else equal?* We therefore construct two synthetic priors that are identical in every respect except the graph generator.

Shared backend. Both priors are implemented inside the same generation pipeline. Given a DAG G on N nodes with a valid topological order, root nodes are sampled from a shared input sampler and each non-root node v is assigned a mechanism

$$X_v \;=\; \phi_v\big(W_v\, X_{\mathrm{Pa}(v)} + b_v\big) + \varepsilon_v, \qquad \varepsilon_v \sim \mathcal{N}(0, \sigma^2), \tag{4}$$

where ϕ_v is drawn from a shared activation library, W_v uses a shared initialisation, and values are standardised per node to keep signals bounded along long paths. One downstream node is designated the target y and the remaining non-root nodes furnish X; the continuous (X, y) then pass through the *identical* downstream stack used by the base model (regression-to-classification labelling, categorical conversion, outlier clipping, standardisation, constant-feature removal, train/test splitting). Full details are in Appendix B.1. Crucially, **Eq.** (4)**, the labelling, and all post-processing are identical across the two arms**, and in both the target is a terminal effect node rather than a cause. The two generators realise this slightly differently—the MLP arm samples y among the outputs of the final layer(s), while the BN arm takes the most-downstream node of the sampled DAG—because the notion of "final layer" has no counterpart in a general DAG; Appendix B.1 gives both rules.

The only difference: the DAG generator. In the **MLP arm** (baseline), G is the layered, block-sparsified feed-forward graph induced by a randomly initialised MLP-SCM, exactly as in TabICL. In the **BN arm** (ours), G is an explicit Bayesian network sampled from a mixture of graph families—scale-free, modular, forest, locally-connected, and Erdős–R'enyi Erdős & R'enyi (1960)—whose per-node in-degree distribution and maximum parent count are matched to the empirical profile of $\mathcal{G}_{\mathrm{RW}}$ (Section 5), with the total node budget fixed to equal the MLP arm's, so the two arms differ in *how* N nodes are wired, not in *how many*. The full sampler and mixture weights are in Appendix B.3; a key property, proved in Appendix B.4, is that constructive scale-free generation yields average degree centrality $\mu_{C_D} = 2\bar{k}/N$ and a Markov blanket that is $O(1)$ in N, so realistic sparsity is achieved by construction rather than by rejection sampling—avoiding the density trap that makes post-hoc filtering of Erdős–R'enyi graphs intractable at $N > 30$.

Why this is the right experiment. Because the two arms share generator mechanism, labelling, model architecture, optimiser, learning-rate schedule, data budget, and random seed, any difference in downstream performance is attributable to DAG topology alone. This is a substantially cleaner test of the structural hypothesis than continued pre-training or generator-family swaps, both of which confound topology with other factors. We note one respect in which the arms are not perfectly matched: the BN arm samples from five graph families while the MLP arm has a single mechanism, so it differs from the baseline in generator *heterogeneity* as well as topology. We discuss this in Section 8 and Appendix D.1, and identify a single-family control as the priority follow-up.

7 Experiments

Setup. We train ICL models from random initialisation, one per prior, on 10^6 synthetic datasets each (single epoch, no repetition), using an identical transformer architecture and optimisation recipe; the paired runs differ only in `prior_type` $\in \{\mathrm{MLP}, \mathrm{BN}\}$. To probe the role of feature dimensionality directly—the axis our structural analysis predicts should matter most—we run the ablation at *two* scales, denoted **f30** and **f60**: f30 caps features at 30 and uses a lightweight activation library, while f60 caps features at 60, enlarges the node budget proportionally, and uses the

Table 1: Head-to-head on the augmented suite (219 datasets). Baseline is the from-scratch MLP arm; Δ is BN$-$MLP. Bootstrap 95% CIs on the mean accuracy delta are given in the text. *Small* and *Large* split the suite at 10,000 training rows ($n = 146$ and $n = 73$). The advantage is present in both prior configurations.

Exp.	Prior	Mean Acc	Log-Loss	Small ($\leq$10K)	Large ($>$10K)
f30	MLP (baseline)	0.7962	0.4753	0.7939	0.8022
	BN (ours)	**0.8043**	**0.4650**	**0.8049**	**0.8027**
	Δ (BN$-$MLP)	$+0.0081$	-0.0103	$+0.0110$	$+0.0005$
f60	MLP (baseline)	0.7847	0.5001	0.7862	0.7807
	BN (ours)	**0.7889**	**0.4927**	**0.7911**	**0.7829**
	Δ (BN$-$MLP)	$+0.0042$	-0.0074	$+0.0049$	$+0.0022$

full activation library (both libraries are enumerated in Appendix B.1). Full training and architecture hyperparameters are in Appendix B.7. All comparisons use the step-1800 checkpoint for both arms in both experiments; both arms' training curves have flattened well before this point—cross-entropy and accuracy plateau by $\sim$1500 steps (Appendix B.6)—so step 1800 reflects a converged model rather than an arbitrary cut-off, and holding the step count fixed across arms and experiments removes checkpoint selection as a confound.

Evaluation suite. Both models are evaluated with the base model's standard inference pipeline, with the MLP arm as the reference against which the BN arm is compared. We begin from the public TALENT benchmark Ye et al. (2024) (167 classification datasets after excluding the 15 TabPFNv2 development sets). Because TALENT is sparse in the high-dimensional regime—only $\sim$30 datasets have ≥ 50 features and only 11 have ≥ 100—and because that regime is central to our hypothesis, we augment it under a rule *specified in advance*: from the curated OpenML suites and a high-feature-count metadata sweep we add every classification dataset that is absent from TALENT, has 50–1000 features and ≥ 2 classes, is not a flattened-image or synthetic near-duplicate family, and loads without error, capping each near-duplicate family at two datasets to avoid pseudo-replication (full rule in Appendix B.5). This yields 52 additional real-world datasets (medical, spectroscopy, sensor, molecular, text, and physics domains), for an augmented suite of 219; we report the full suite and, for robustness, the TALENT-only and external-only subsets separately.

Training dynamics. Both arms train stably to convergence, and on its own prior the BN arm reaches lower training cross-entropy and higher training accuracy (Appendix B.6). Because each arm is measured on its *own* prior this is not decisive alone, but it is consistent with the transfer results below: a merely easier-to-fit prior would show the former without the latter.

Aggregate result. Table 1 summarises head-to-head performance on the augmented suite; in both experiments the BN prior improves mean accuracy and log-loss over the MLP prior. A paired bootstrap (10^4 resamples) over the 219 per-dataset accuracy deltas gives, for f30, a mean of $+0.81\%$ with 95% CI $[+0.38\%, +1.28\%]$ and $P(\text{mean} \leq 0) < 10^{-3}$, and for f60 a mean of $+0.42\%$ with 95% CI $[-0.06\%, +0.89\%]$ and $P(\text{mean} \leq 0) = 0.041$. The two-sided Wilcoxon signed-rank test on the median is not significant in either experiment: the BN prior does not win on the *typical* dataset, but its wins exceed its losses in magnitude. Both arms are also absolutely weaker at f60 than at f30 (MLP $0.7962 \rightarrow 0.7847$) despite the larger node budget and richer activation library; the step budget is held fixed across experiments, so the f60 models see equally many but harder datasets and are undertrained relative to task difficulty, which does not affect the within-experiment contrast. **The BN advantage is real but not uniform—it is concentrated, and (as we show next) that concentration is in the high-dimensional regime.**

The advantage concentrates in, and grows with, feature dimensionality. Figure 2(a) resolves the aggregate into feature-count buckets for both experiments. A consistent pattern emerges. In the low- and mid-dimensional range the advantage is small (and, at 21–50 features under f30, significantly negative—see below), but in the high-dimensional range the BN prior wins significantly,

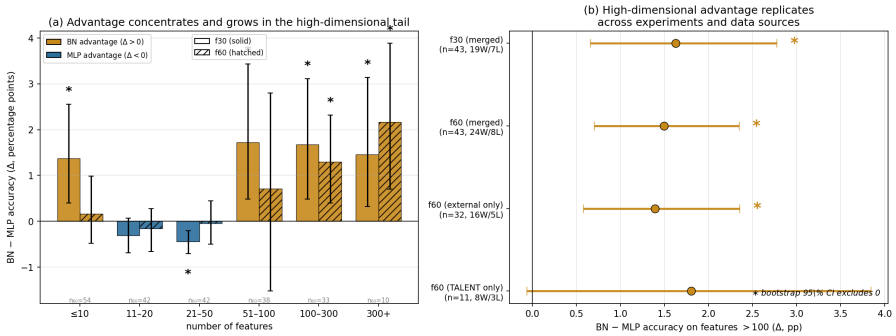


Figure 2: **(a)** BN−MLP accuracy delta by feature-count bucket for the f30 (solid) and f60 (hatched) priors; error bars are bootstrap 95% CIs, $*$ marks CIs excluding zero. The advantage is small or negative in the low/mid range and significant in the high-dimensional range, growing to $+2.2$pp at > 300 features (f60). Orange: BN advantage ($\Delta > 0$); blue: MLP ($\Delta < 0$). **(b)** The > 100-feature advantage across independent slices: CI-significant for f30 and f60 on the merged suite and on the external-only datasets (disjoint from TALENT), and same-signed on the underpowered TALENT-only slice—replicating across prior configurations and data sources.

and the effect *grows in the tail*: under f60 the mean delta rises monotonically from $+0.7\%$ (51–100 features) to $+1.3\%$ (101–300) to $+2.2\%$ (> 300), with the latter two bootstrap-CI-significant despite small counts. Aggregating all datasets with > 100 features, the BN prior improves accuracy by $+1.63\%$ under f30 (bootstrap CI $[+0.65\%, +2.78\%]$; Wilcoxon $p = 0.0019$) and $+1.50\%$ under f60 (CI $[+0.70\%, +2.34\%]$; Wilcoxon $p = 0.0020$); both survive Holm–Bonferroni correction across all tested subgroups ($p < 0.05$). This is the paper's central quantitative result. Because f30 and f60 also differ in node budget and activation library, we read the tail growth *within* each experiment (across feature-count buckets), rather than the difference *between* experiments, as the evidence that the advantage grows with dimensionality.

The high-dimensional effect is robust, not a benchmark artifact. Because the augmentation deliberately enriches the high-dimensional regime, a natural concern is that we merely added datasets that happen to favour the BN prior. Figure 2(b) rules this out: the > 100-feature advantage is CI-significant not only on the merged suite in both experiments but on the *external OpenML datasets alone* ($n = 32$, disjoint from TALENT): $+1.39\%$, CI $[+0.58\%, +2.35\%]$, Wilcoxon $p = 0.005$, Holm-significant. The TALENT-only slice points the same way ($+1.80\%$) but is underpowered at $n = 11$. Replication across two independently trained prior configurations *and* two independent dataset sources is strong evidence that the effect is a property of the prior rather than of any particular benchmark.

An honest negative, and a fragile positive. Two secondary findings delimit the claim. In the f30 experiment the mid-dimensional bucket (21–50 features) is the one regime where the BN prior is significantly *worse* (-0.44%, Holm-corrected $p < 0.01$): the mid-range is where the MLP prior's block-structured dependence is the closest match, so a uniform "BN is better" claim would be false. And a striking low-dimensional effect from our f30 pilot—the BN prior winning all nine financial (FOREX) datasets by $+7.6\%$—*does not survive* the richer f60 prior ($4/9$, not significant). Domain-specific low-dimensional gains can therefore be real but fragile to prior design, whereas the high-dimensional advantage is robust across configurations and sources: the effect of topology is *regime-specific*, and the dependable regime is the high-dimensional one.

8 Discussion and Limitations

Concentrated, not uniform. The core prediction of our structural analysis (Section 5.2)—that the BN prior's benefit should be largest where the MLP prior's conditional width scales worst, i.e. at high feature dimensionality—is borne out in both experiments, grows monotonically in the feature-count tail, and replicates on external data. Because f30 and f60 differ in node budget and activation library as well as feature cap, they are best read as two *independent configurations* of the same hypothesis

rather than a dimensionality-scaling curve; that the effect survives both is what makes it credible, and a single-variable scaling study is natural future work. The aggregate effect is CI-significant on the mean but never median-significant, so we do not overclaim: topology is a lever whose benefit is concentrated in the high-dimensional regime, small elsewhere, and occasionally negative in the dense mid-range.

Extrapolation beyond the training feature cap. Our models are trained with at most 30 and 60 features while the headline result is measured on tables with more than 100 (up to 970), so the advantage could reflect either genuine competence on wide, modular dependence structures or merely gentler degradation outside the training range. We separate these by truncating each of the 43 high-dimensional datasets to 60 and to 30 features—inside both training caps, so that no extrapolation remains—under two selection rules: retaining the features with highest train-split mutual information with the target (TOP-k, which preserves the informative dependence structure), and retaining a random subset (three seeds, which does not). Width extrapolation is not the explanation. At 30 TOP-k features the f60 BN advantage is $+0.86$pp (CI $[+0.34, +1.45]$, Wilcoxon $p = 0.005$)—significant on the median as well as the mean, despite every table now lying inside the training feature cap. Moreover, at matched width the advantage is roughly twice as large when structure is preserved as when it is not ($+1.24$ vs. $+0.60$pp at 60 features under f60; $+0.68$ vs. $+0.37$ under f30), which is direct evidence for the structural reading. The advantage does attenuate monotonically as features are removed ($+1.50 \rightarrow +1.24 \rightarrow +0.86$pp under f60), so part of it is specific to the wide regime and both mechanisms contribute. Full results, and the caveat that TOP-k selection also removes noise features and so does not isolate structure preservation from denoising, are in Appendix D.4.

Three further limitations. We record three caveats here and expand them in Appendix D. *First*, the BN arm differs from the MLP arm in generator heterogeneity (five graph families versus one) as well as in topology, and the two are not separable in the present design; we argue in Appendix D.1 that the topological reading is the more likely explanation—the advantage is regime-specific and even negative in the mid-range, whereas a diversity benefit should help uniformly—and identify a single-family BN arm as the priority follow-up. *Second*, both arms are trained at reduced scale and trail the published TabICL checkpoint absolutely; we do not claim our deltas are a lower bound on full scale, since a larger model might learn to discount the MLP prior's wide conditionals (Appendix D.2). *Third*, because the augmentation enriches the high-dimensional regime, the aggregate could be biased; the safeguards—rule fixed in advance, subsets reported separately, and the claim resting on the bucketed and external-only results rather than the aggregate mean—are detailed in Appendix D.3. We note that the rule was fixed internally in advance but not lodged as a public timestamped protocol, so it does not constitute pre-registration in the formal sense.

9 Conclusion

We recast the design of synthetic priors for tabular foundation models as a question of *graph topology*, showed that MLP-induced priors diverge systematically from real causal graphs—presenting conditional dependence structures an order of magnitude wider, and widening further as graphs grow—and tested the consequences under a clean from-scratch ablation that varies topology alone. Across two prior configurations and an augmented suite of 219 real datasets, real-world-matched Bayesian network topology yields a benefit that is concentrated in, and grows with, feature dimensionality: on tables with more than 100 features it improves accuracy by $+1.5$–1.6 percentage points, a result that survives multiple-comparison correction, replicates on external datasets disjoint from the base benchmark, and persists when those tables are truncated to the training feature width provided the retained features are informative—so it is not an artifact of extrapolating beyond the widths seen in training. Low- and mid-dimensional effects are smaller or fragile. DAG topology is thus a genuine, under-explored axis of prior design, with a benefit that is sharpest precisely where causal-structural theory predicts the discrepancy between MLP-like and real-network-like priors to be largest. We offer this as a guideline: practitioners building the next generation of tabular foundation models should treat the topology of the synthetic prior as a tunable, consequential design choice, and should test structure-aligned priors at full scale.

REFERENCES

South German Credit. UCI Machine Learning Repository, 2020. DOI: https://doi.org/10.24432/C5QG88.

Ankur Ankan and Johannes Textor. pgmpy: A Python toolkit for Bayesian networks. *Journal of Machine Learning Research*, 25(265):1–8, 2024. URL https://jmlr.org/papers/v25/23-0487.html.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Paulo Cortez. Student Performance. UCI Machine Learning Repository, 2008. DOI: https://doi.org/10.24432/C5TG7T.

Felix den Breejen, Sangmin Bae, Stephen Cha, and Se-Young Yun. Fine-tuned in-context learning transformers are excellent tabular data classifiers. *arXiv preprint arXiv:2405.13396*, 2024.

Paul Erdős and Alfr'ed R'enyi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1):17–60, 1960.

Francis Galton. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, Dominik Safaric, Jake Robertson, Benjamin Jäger, Simone Alessi, Adrian Hayler, Vladyslav Moroshan, Lennart Purucker, Philipp Singer, Alan Arazi, Julien Siems, Jan Hendrik Metzen, Georg Grab, Nick Erickson, Siyuan Guo, Eliott Kalfon, Simon Bing, David Salinas, Clara Cornu, Lilly Charlotte Wehrhahn, Diana Kriuchkova, Kursat Kaya, Lydia Sidhoum, Marie Salmon, Jerry Chen, Madelon Hulsebos, Yann LeCun, Samuel Müller, Bernhard Schölkopf, Sauraj Gambhir, Noah Hollmann, and Frank Hutter. Tabpfn-3: Technical report, 2026. URL https://arxiv.org/abs/2605.13986.

David E. Heckerman, Eric J. Horvitz, and Bharat N. Nathwani. Toward normative expert systems: Part I. the pathfinder project. *Methods of Information in Medicine*, 31(2):90–105, 1992.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second, 2023. URL https://arxiv.org/abs/2207.01848.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Mads Ipsen and Alexander S. Mikhailov. Evolutionary reconstruction of networks. *Physical Review E*, 66(4):046109, 2002.

Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, 2013.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Jesse C Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L Caterini, and Maksims Volkovs. Tabdpt: Scaling tabular foundation models on real data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Alessandro Magrini, Silvia Di Blasi, and Federico M. Stefanini. A conditional linear gaussian network to assess the impact of several agronomic settings on the quality of tuscan sangiovese grapes. *Biometrical Letters*, 54(1):25–42, 2017.

Stefan McCabe, Leo Torres, Timothy LaRock, Syed Arefinul Haque, Chia-Hung Yang, Harrison Hartle, and Brennan Klein. netrd: A library for network reconstruction and graph distances. *Journal of Open Source Software*, 6(62):2990, 2021.

National Institute of Diabetes. Pima indians diabetes database. UCI Machine Learning Repository, 1990.

Rainer Opgen-Rhein and Korbinian Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 1(37), 2007. doi: 10.1186/1752-0509-1-37.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. Tabicl: A tabular foundation model for in-context learning on large data. In *International Conference on Machine Learning*, pp. 50817–50847. PMLR, 2025.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

J. R. Quinlan. Credit Approval. UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C5FS30.

R. Quinlan. Auto MPG. UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5859H.

Yoram Reich and Steven Fenves. Pittsburgh Bridges. UCI Machine Learning Repository, 1989. DOI: https://doi.org/10.24432/C5RP5H.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721): 523–529, 2005.

V. E. Sathishkumar and Yongyun Cho. A rule-based model for seoul bike sharing demand prediction using weather data. *European Journal of Remote Sensing*, 53(1):166–183, 2020. doi: 10.17632/zbdtzxcxvg.2.

Marco Scutari. Learning Bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3):1–22, 2010. doi: 10.18637/jss.v035.i03.

Marco Scutari. Bayesian networks, MAGIC populations and multiple trait prediction. Invited talk, 5th International Conference on Quantitative Genetics (ICQG 2016). Network distributed via the `bnlearn` Bayesian Network Repository, 2016.

Marco Scutari, Phil Howell, David J. Balding, and Ian Mackay. Multiple quantitative trait analysis using bayesian networks. *Genetics*, 198(1):129–137, 2014.

William H Sewell and Vimal P Shah. Social class, parental encouragement, and educational aspirations. *American Journal of Sociology*, 73(5):559–572, 1968.

Jin Tian, Azaria Paz, and Judea Pearl. Finding minimal d-separators. Technical Report R-254, Cognitive Systems Laboratory, UCLA, 1998.

Claudia Vitolo, Marco Scutari, Mohamed Ghalaieny, Allan Tucker, and Andrew Russell. Modelling air pollution, climate and health data using bayesian networks: a case study of the english regions. *Earth and Space Science*, 5(4):76–88, 2018.

Jonas Wahl and Jakob Runge. Separation-based distance measures for causal graphs, 2025. URL `https://arxiv.org/abs/2402.04952`.

Han-Jia Ye, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and De-Chuan Zhan. A closer look at deep learning methods on tabular datasets. *arXiv preprint arXiv:2407.00956*, 2024. URL `https://arxiv.org/abs/2407.00956`.

Xiyuan Zhang, Danielle C. Maddix, Junming Yin, Nick Erickson, Abdul Fatir Ansari, Boran Han, Shuai Zhang, Leman Akoglu, Christos Faloutsos, Michael W. Mahoney, Cuixiong Hu, Huzefa Rangwala, George Karypis, and Bernie Wang. Mitra: Mixed synthetic priors for enhancing tabular foundation models, 2025. URL `https://arxiv.org/abs/2510.21204`.

A Background on Causal Graphs: Full Details

This appendix expands the definitions used informally in Section 3 and gives the derivations underlying the metrics of Section 5.1.

A.1 Notation and the causal Markov condition

A *causal graph* is a directed acyclic graph (DAG) $G = (V, E)$ whose vertices $V = \{X_1, \dots, X_N\}$ are random variables and whose directed edges $X_i \to X_j$ denote a direct causal influence of X_i on X_j. We write $\mathrm{Pa}(X_i)$ for the parents of X_i, $\mathrm{Ch}(X_i)$ for its children, $\mathrm{An}(X_i)$ for its ancestors, and $\mathrm{De}(X_i)$ for its descendants; for a set S we write $\mathrm{An}(S) = \bigcup_{v \in S} \mathrm{An}(v) \cup S$. Under the Causal Markov Condition the joint factorises as in Eq. (1), i.e. every variable is independent of its non-descendants given its parents. The *Markov blanket* $\mathrm{MB}(X_i) = \mathrm{Pa}(X_i) \cup \mathrm{Ch}(X_i) \cup \{\mathrm{Pa}(c) : c \in \mathrm{Ch}(X_i)\}$ is the minimal set that renders X_i conditionally independent of all other variables. The sample complexity of learning $P(X_i \mid \mathrm{MB}(X_i))$ grows with $|\mathrm{MB}(X_i)|$—for discrete variables with s states the conditional has $O(s^{|\mathrm{MB}(X_i)|})$ free parameters—so sparse, bounded-parent graphs are dramatically more sample-efficient to learn than wide ones. This is the observation that motivates the entire paper, and the reason blanket width rather than any N-normalised density is our primary structural statistic.

A.2 Path configurations and d-separation

Association flows along paths according to three elementary configurations at an intermediate node Z on a path between X and Y:

- **Chain** $X \to Z \to Y$: transmits association; blocked by conditioning on Z. Gives $X \perp\!\!\!\perp Y \mid Z$.
- **Fork** $X \leftarrow Z \to Y$: transmits association; blocked by conditioning on Z. Gives $X \perp\!\!\!\perp Y \mid Z$.
- **Collider** $X \to Z \leftarrow Y$: blocks association when Z (and all its descendants) are *not* conditioned on; conditioning on Z or a descendant *opens* the path (Berkson's paradox). Gives $X \perp\!\!\!\perp Y$ marginally but $X \not\perp\!\!\!\perp Y \mid Z$.

A path is *blocked* by a conditioning set $\mathbf{Z}$ if it contains a chain or fork whose middle node is in $\mathbf{Z}$, or a collider whose middle node and all of whose descendants are outside $\mathbf{Z}$. Two nodes u, v are *d-separated by* $\mathbf{Z}$, written $u \perp_d v \mid \mathbf{Z}$, if $\mathbf{Z}$ blocks *every* path between them. D-separation is the graphical criterion for conditional independence: $u \perp_d v \mid \mathbf{Z}$ in G implies $X_u \perp\!\!\!\perp X_v \mid \mathbf{Z}$ under every distribution Markov to G Pearl (2009).

A.3 Markov blankets, minimum d-separators, and the separator bound

Markov blanket as a moral neighbourhood. Given $G = (V, E)$, the *moral graph* $M = (V, E_M)$ is formed by adding an undirected edge between every pair of co-parents of a common child and then dropping all edge directions. By construction the neighbourhood of v in M is exactly $\mathrm{Pa}(v) \cup \mathrm{Ch}(v) \cup \{\mathrm{Pa}(c) : c \in \mathrm{Ch}(v)\} = \mathrm{MB}(v)$, so

$$|\mathrm{MB}(v)| \;=\; \deg_M(v), \qquad \text{and} \qquad \frac{1}{N}\sum_v |\mathrm{MB}(v)| \;=\; \frac{2|E_M|}{N}. \tag{5}$$

The entire blanket distribution of a graph therefore follows from a single moralisation pass in $O\big(\sum_v |\mathrm{Pa}(v)|^2\big)$ time. This is how the blanket statistics in Figure 1 are computed—exactly, deterministically, on every graph in both corpora, with no sampling and no exclusions.

Minimum d-separators via ancestral moralisation. For a non-adjacent pair (u, v), let $A = \mathrm{An}(\{u, v\})$ and let $M_A = \mathrm{moral}(G[A])$. Tian et al. (1998) show that a minimum-cardinality d-separator of u and v is a minimum u–v vertex cut in M_A, so

$$|\mathbf{Z}_{\min}(u, v)| \;=\; \kappa_{M_A}(u, v), \tag{6}$$

the local vertex connectivity of u and v in M_A, computable as a maximum-flow / minimum-vertex-cut. The restriction to the ancestral subgraph cannot be dropped. If u and v are non-adjacent in G they can never be adjacent in M_A: a moral edge $u - v$ would require a common child $w \in A$, but $w \in \mathrm{De}(u) \cap \mathrm{An}(\{u, v\})$ implies a directed cycle. Moralising the *entire* DAG, by contrast, marries co-parents everywhere and so creates precisely the adjacencies the correct construction excludes, declaring inseparable many pairs that are in fact d-separated by the empty set. We also note that $\kappa_{M_A}(u, v) = 0$ is meaningful and common in large sparse graphs: it indicates a pair already d-separated by $\emptyset$, i.e. marginally independent. Because of this the per-graph *maximum* separator cardinality is the more robust statistic than the mean, being insensitive to the fraction of trivially independent sampled pairs; we report both.

Proof of Theorem 1. Let u, v be non-adjacent in G. Since G is acyclic, at most one of the two is a descendant of the other; assume without loss of generality $v \notin \mathrm{De}(u)$. As u and v are non-adjacent, $v \notin \mathrm{Pa}(u)$, so v is a non-descendant of u lying outside $\mathrm{Pa}(u)$. The local Markov property gives $u \perp_d v \mid \mathrm{Pa}(u)$, so $\mathrm{Pa}(u)$ is an admissible separator and $|\mathbf{Z}_{\min}(u, v)| \le |\mathrm{Pa}(u)| \le \max_{w \in V} |\mathrm{Pa}(w)|$. □

A corollary, due originally to Verma and Pearl, is that *every* non-adjacent pair in a DAG is d-separable by some set. We verified Theorem 1 empirically as a correctness check on our implementation: across all 3,033 graphs for which we compute both quantities there are zero violations of $\max_{u,v} |\mathbf{Z}_{\min}(u, v)| \le \max_w |\mathrm{Pa}(w)|$.

What we do not report, and why. Let $S(G)$ denote the number of d-separable non-adjacent pairs of G. By the corollary above every non-adjacent pair is d-separable, and adjacent pairs admit no separator, so

$$S(G) \;=\; \binom{N}{2} - |E|, \qquad \frac{S(G)}{\binom{N}{2}} \;=\; 1 - \text{density}(G), \tag{7}$$

exactly. Counts of separable pairs are therefore a closed-form function of size and density, carrying no topological information beyond what the node-count and density panels already display; and at fixed N a *denser* graph has *fewer* such pairs, so the quantity moves opposite to the direction a density argument would require. In our corpora the normalised version of Eq. (7) is not significant at the 99% level ($p = 0.031$; TabICL 0.936 vs. real-world 0.905). We record this explicitly so that the reader can see why the comparison in this paper is formulated in terms of separator *cardinality* (Eq. (6)) and blanket *width* (Eq. (5)) rather than pair counts.

A.4 Structural metrics: definitions

Let $G = (V, E)$ with $N = |V|$. Unless stated otherwise all quantities are computed on the *directed* graph, using `networkx` defaults.

Markov-blanket width. $|\mathrm{MB}(v)| = \deg_M(v)$ as in Eq. (5). We report the per-graph mean, median, and maximum. We deliberately do *not* normalise by N: the conditional $P(X_v \mid \mathrm{MB}(v))$ has dimension $|\mathrm{MB}(v)|$, so it is the absolute width that governs sample complexity, and dividing by N reintroduces exactly the $1/N$ artifact that inverts the density and degree-centrality comparisons. For completeness, the normalised quantity is nearly equal across corpora (median $|\mathrm{MB}|/N$: 0.165 for TabICL vs. 0.150 real-world), which is why the *size-invariance* of real-world width (Figure 1a), rather than the ratio, is the evidence that the gap is topological.

Minimum d-separator cardinality. $|\mathbf{Z}_{\min}(u, v)| = \kappa_{M_A}(u, v)$ as in Eq. (6), computed with `networkx` local node connectivity. This is the minimum-*cardinality* separator, not a *minimal* (irreducible) one; the two differ, and only the former is comparable across graphs. Non-adjacent pairs are sampled uniformly at random: exhaustively for the real-world corpus, and 150 pairs on each of 3,000 uniformly sampled SCMs. We report the per-graph mean, median, and maximum.

Degree centrality and mean parent count.

$$C_D(v) \;=\; \frac{\deg(v)}{N - 1}, \tag{8}$$

where $\deg(v)$ counts incident edges irrespective of orientation. Since $\sum_v \deg(v) = 2|E|$, the graph-level average is $\mu_{C_D} = 2|E|/\big(N(N-1)\big)$, which is identical up to a factor $N/(N-1)$ to the normalised density and is precisely the quantity analysed in Eq. (11). The $1/(N-1)$ normalisation is what makes μ_{C_D} decay as $2\bar{k}/N$ under a linear edge budget—and what makes it an unsuitable proxy for conditional width. By *mean parent count* we mean instead $|E|/N = \frac{1}{N}\sum_v |\mathrm{Pa}(v)|$, the unnormalised in-degree average; this is the quantity we intend by "density" throughout the paper, because it is the per-node measure appearing in the Markov factorisation (1) and it carries no $1/N$ factor.

Closeness centrality. For graphs in which not every pair is mutually reachable—the generic case for a DAG—we use the Wasserman–Faust correction,

$$C_C(v) \;=\; \frac{n_v - 1}{N-1} \cdot \frac{n_v - 1}{\sum_{u \in R(v)} d(u,v)}, \tag{9}$$

where $R(v)$ is the set of nodes from which v is reachable, $n_v = |R(v)|$, and $d(u,v)$ is the shortest-path length. Source nodes, which no other node can reach, receive $C_C(v) = 0$; because both corpora contain such nodes by construction, the *minimum* closeness is identically zero in every graph and carries no information, so it is not reported.

Betweenness centrality.

$$C_B(v) \;=\; \frac{1}{(N-1)(N-2)} \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}, \tag{10}$$

where σ_{st} is the number of shortest directed s–t paths and $\sigma_{st}(v)$ the number of those passing through v. The $1/\big((N-1)(N-2)\big)$ factor normalises by the number of ordered node pairs, so values are comparable between a 20-node and a 1041-node graph.

Intra-set structural dissimilarity. To characterise the internal diversity of a corpus we use the Ipsen–Mikhailov index Ipsen & Mikhailov (2002) as implemented in `netrd` McCabe et al. (2021) at its default kernel half-width. In its original undirected formulation one writes $L = D - A$, takes vibrational frequencies $\omega_i = \sqrt{\lambda_i}$ from the Laplacian eigenvalues, smooths them into a density ρ_G with a Lorentz kernel, and defines $\varepsilon(G,H) = \big(\int_0^\infty [\rho_G - \rho_H]^2 \, d\omega\big)^{1/2}$. We pass the directed adjacency matrices without symmetrisation, and we verified that the implementation is orientation-sensitive (for a DAG g, $\varepsilon(g, g^{\mathrm{rev}}) = 0.26 \neq 0$), so the quantity we obtain is *not* the undirected Laplacian spectral distance. Absolute values are also implementation-dependent, since libraries differ in how the kernel half-width is fixed. We therefore treat ε as a *reproducible structural dissimilarity index* rather than a spectral metric: it is computed identically for both corpora, so between-corpus comparisons are internally consistent. For each graph g in a corpus $\mathcal{C}$ we compute $\varepsilon(g,h)$ for all $h \in \mathcal{C} \setminus \{g\}$ and summarise by the mean and minimum; a corpus of near-duplicates yields small values, a heterogeneous corpus large ones. Because the between-corpus comparison does not survive matching on corpus size and node count (Appendix A.6), we draw no conclusion from it in the main text.

A.5 Supplementary structural metrics

Figure 3 reports the remaining comparisons. The top and middle rows show six *significant* metrics. Mean Markov-blanket width (43.8 vs. 4.0, $r_{rb} = 0.979$, $p < 10^{-21}$) and mean parent count $|E|/N$ (8.50 vs. 1.52, $r_{rb} = 0.962$, $p < 10^{-20}$) are the two panels displaced from the main figure; they carry the largest effect sizes in the entire comparison. Node counts differ by $9\times$ ($p < 10^{-13}$), maximum Markov-blanket width by $6.6\times$ (86 vs. 13, $p < 10^{-16}$), and mean minimum-d-separator cardinality by $4.6\times$ (5.05 vs. 1.09, $p < 10^{-13}$); maximum degree centrality is higher for real-world graphs ($p = 1.1 \times 10^{-4}$), corroborating hub concentration.

The bottom row shows metrics that do *not* clear the 99% threshold. Normalised density and average degree centrality ($p = 0.031$ for both; they are proportional by Eq. (8)) are *higher* for real-world graphs, and average closeness centrality is statistically indistinguishable ($p = 0.94$; maximum closeness likewise, $p = 0.87$). As explained in Section 5, the density and degree direction is an

arithmetic consequence of the $1/N$ factor in both normalisations rather than a contradiction of the per-node findings: real-world graphs have $5.6\times$ fewer parents per node while being $9\times$ smaller, and $2|E|/(N(N-1))$ rewards small size. We rest no claim on any metric in this row.

Sensitivity to the small learned networks. Ten real-world graphs have fewer than 20 nodes. Removing all of them leaves the gaps essentially unchanged: mean blanket width $10.9\times \rightarrow 10.0\times$, mean parent count $5.61\times \rightarrow 5.67\times$, maximum separator cardinality $6.67\times \rightarrow 6.67\times$, maximum blanket width $6.6\times \rightarrow 3.9\times$, and maximum parent count $10.3\times \rightarrow 7.8\times$ (all $p < 10^{-10}$). Within the matched band $N \in [100, 500]$—containing six real-world graphs (ANDES, ARTH150, DIABETES, MUNIN1, PATHFINDER, PIGS)—the blanket-width gap is $10.2\times$ ($r_{rb} = 0.999$, $p = 2.3 \times 10^{-5}$). No conclusion in this paper depends on the small networks or on the size difference between the corpora.

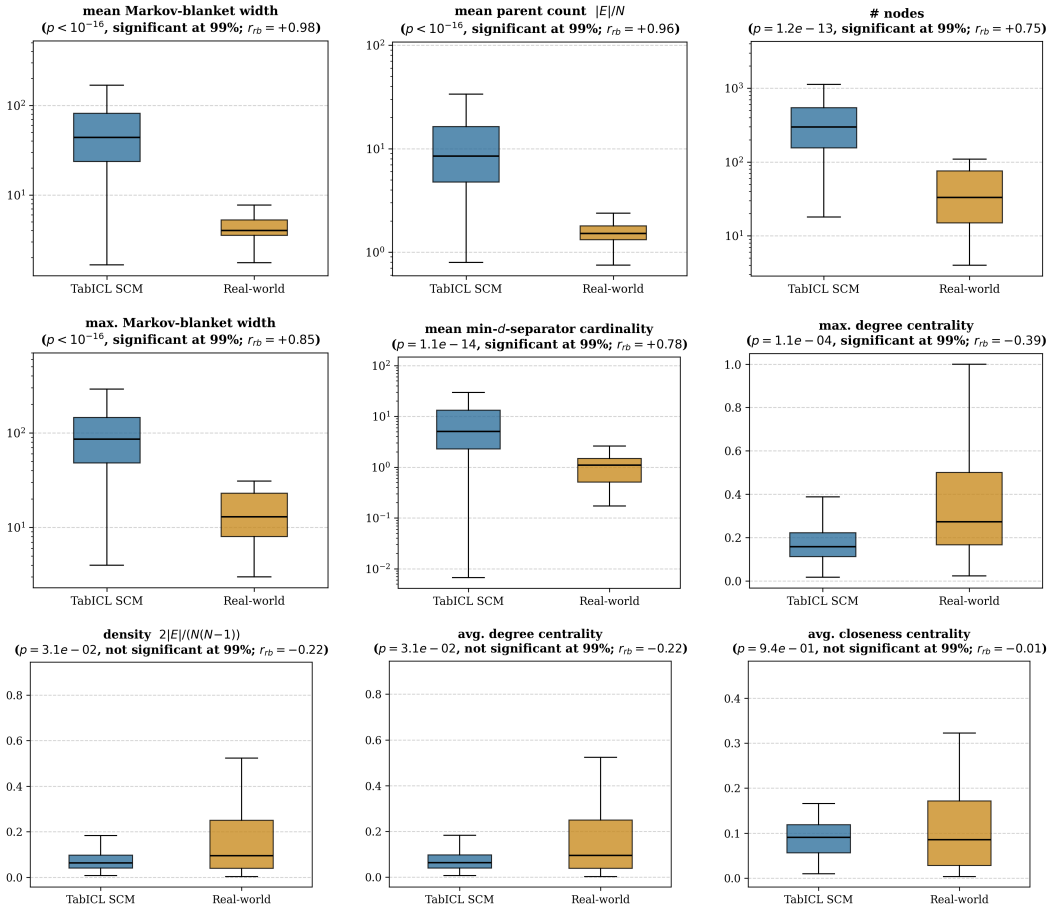


Figure 3: Supplementary structural metrics. **Top and middle rows (significant at 99%):** mean Markov-blanket width, mean parent count $|E|/N$, node count, maximum Markov-blanket width, mean minimum-d-separator cardinality, and maximum degree centrality. The first two carry the largest effect sizes in the comparison ($r_{rb} = 0.979$ and 0.962) and are displaced here only for space. **Bottom row (not significant):** normalised density, average degree centrality, and average closeness centrality; the first two are proportional and both run *opposite* to the per-node measures, for the arithmetic reason given in Section 5. Significance by Mann–Whitney U; p and rank-biserial r_{rb} annotated on each panel.

A.6 STRUCTURAL DIVERSITY: A NEGATIVE CONTROL

We had expected the randomly sampled MLP corpus to be structurally *redundant* relative to the heterogeneous real-world corpus, and the raw intra-set dissimilarity indices are consistent with that

(mean 0.628 vs. 0.693; minimum 0.152 vs. 0.350). Intra-set dissimilarity is, however, confounded in two ways that both favour the hypothesis: each SCM has $1{,}199$ candidate neighbours against 32 for a real-world graph, which mechanically lowers its *nearest*-neighbour distance; and the corpora differ in node count, to which the index is sensitive. Table 2 reports both controls. The minimum-distance gap is largely an artifact of neighbour count ($0.152 \to 0.279$ once the SCM corpus is subsampled to 33 graphs, closing $\sim 77\%$ of the gap). The mean-distance gap survives in direction but narrows, and only 61% of node-matched draws clear $p < 0.01$. We therefore report the diversity comparison as directionally consistent but *not robust*, and no claim in the paper rests on it.

Table 2: Intra-set structural dissimilarity under two confound controls (medians; 200 resampling draws for the matched rows). *Neighbour-matched* subsamples $\mathcal{G}_{\mathrm{SCM}}$ to 33 graphs, equalising the number of candidate neighbours. *Node-matched* additionally draws SCMs of comparable node count (median $|\Delta N| = 1$). The final column is a two-sided Mann–Whitney test on the mean index.

Control	SCM mean	RW mean	SCM min	RW min	n_{SCM}	p (mean)
Unmatched	0.628	0.693	0.152	0.350	1200	< 0.001
Neighbour-matched	0.626	0.693	0.279	0.350	33	0.002
Node-matched	0.612	0.693	0.281	0.350	21	0.002

A.7 A mechanistic hypothesis: conditional width as causal noise

We offer an interpretation, consistent with but not proven by our results, for *why* topology matters most at high dimensionality. MLP priors present wide conditionals whose width grows with the size of the graph (Section 5), so the induced label function is high-complexity relative to the in-context sample budget. High-dimensional real tables are typically *modular and hub-organised* rather than uniformly wide; the MLP prior cannot represent this efficiently and effectively injects *causal noise*—the transformer must infer a coherent low-order rule from a prior that never presented one—whereas the BN prior's sparse, hub-and-spoke topology matches such structure and presents clean low-order dependencies. The mismatch, and hence the BN advantage, grows with the number of interacting variables, which is the tail growth observed in Figure 2(a). In the dense mid-dimensional range, where the MLP prior's width most closely matches the target, it retains a slight edge—the honest negative of Section 7. The feature-truncation control of Appendix D.4 bears this out: at matched width, the BN advantage is roughly twice as large when the informative dependence structure is retained as when features are chosen arbitrarily.

B Dataset Generation and Evaluation: Full Details

B.1 Shared generation backend

The generation pipeline (Section 6) samples a DAG, samples root values, propagates the shared mechanism of Eq. (4) in topological order, designates one downstream node as the target, and passes the continuous (X, y) through the base model's regression-to-classification stack. Here we record the implementation details omitted from the main text.

Root sampler. Nodes with no parents are drawn from a shared input sampler supporting normal, uniform, and mixed (normal/multinomial/Zipf) marginals, matching the base model's root distribution.

Activation library. The mechanism nonlinearity ϕ_v in Eq. (4) is drawn from an activation library that differs between the two experiments. The **f60** experiment uses the base model's full library. Its base set comprises ten standard elementwise nonlinearities (identity, $\tanh$, ReLU, ReLU6, LeakyReLU, ELU, SELU, SiLU, softplus, hardtanh) and eight deliberately irregular ones: a sign function, a Gaussian radial basis e^{-x^2}, an indicator $\mathbf{1}[|x| \le 1]$, $\sin x$, x^2, $|x|$, $\sqrt{|x|}$, and e^x, together with a random Fourier-feature activation summing 256 sine components with random frequencies and phases whose weights decay as a power law with exponent drawn log-uniformly from $[0.7, 3.0]$. The Fourier activation enters with multiplicity 10, so roughly a third of draws are irregular by construction. Every entry is then wrapped in a composition that standardises its input (statistics fitted on the first batch) and applies a random affine rescaling $x \mapsto s(x + b)$ with $\log s \sim \mathcal{N}(0, 2^2)$ and

$b \sim \mathcal{N}(0, 1)$; the library is finally doubled by adding, for each entry, a variant that re-samples the activation independently at each node, so different nodes of one graph may carry different mechanisms. This yields 56 library entries. The **f30** experiment instead draws uniformly from eight bounded or piecewise-linear activations applied directly, without the scaling wrappers: $\tanh$, ReLU, GELU, SiLU, ELU, LeakyReLU, sigmoid, and identity. This set was chosen for generation throughput—the full library's Fourier activation dominates sampling cost, and restricting to the cheap set reduced BN generation from $\sim$35 to $\sim$16 seconds per batch, which is what made a 10^6-dataset pool tractable under our compute budget. The two sets overlap but neither contains the other. In both experiments *both arms draw from the identical set*, so the activation library is never confounded with the topological variable; it is one of the three respects in which f30 and f60 differ, which is why we read them as two independent configurations rather than a scaling curve (Section 8).

Target selection. Both arms designate a terminal effect node as y (`y_is_effect = True`) and draw the features X uniformly at random without replacement from the remaining non-root nodes, sorted by index. The arms differ in how "terminal" is operationalised, because a layered MLP has a distinguished final layer whereas a general DAG does not: the MLP arm samples y among the outputs of the final layer(s), while the BN arm takes the last node in the topological order. Since the BN sampler introduces nodes in index order, that is the most recently added node; its ancestry is large under the scale-free and uniform families but can be comparatively shallow under the forest and locally-connected families. We note this asymmetry for completeness; it is a consequence of the topological difference under study rather than an independent design choice, and it does not change the fact that in both arms y is an effect of the sampled mechanism.

Numerical guards. The linear map $W_v \in \mathbb{R}^{1\times|\mathrm{Pa}(v)|}$ is initialised with variance $\propto 1/|\mathrm{Pa}(v)|$; to keep signals bounded along long paths, each node's output is clamped to a finite range and standardised to zero mean and unit variance before being passed to its children, and a final `nan_to_num` guard removes any residual non-finite values produced by unbounded activations. These guards are necessary rather than precautionary under the f60 library, whose unbounded members—notably e^x and x^2 under a random affine rescaling with $\log s \sim \mathcal{N}(0, 2^2)$—overflow to non-finite values when composed along the long directed paths that sparse DAGs produce.

Regression-to-classification. The continuous target is standardised and thresholded into a categorical label (rank- or value-based, with random label permutation); a random subset of features is converted to categorical; features undergo outlier clipping, standardisation, and optional permutation; constant features are removed; and the rows are split into an in-context (train) portion and a query (test) portion. Only the DAG-sampling step differs between the two arms; the two experiments additionally differ in the feature cap (30 for f30, 60 for f60), the corresponding node budget, and the activation library, with both arms sharing these settings within each experiment.

B.2 MLP ARM DAG

The MLP arm reproduces the TabICL generator. Root "causes" of count n_{c} feed a feed-forward stack of L layers of width H; within each layer, weights are initialised densely and then sparsified by block-wise dropout (or, equivalently, edge pruning). The induced DAG has edges only between adjacent layers, with no skip or residual connections. The structural hyperparameters (n_{c}, L, H, dropout rate, activation family, noise scale) are sampled from the base model's meta-distributions so that the MLP arm is a faithful instantiation of the published prior; the scaling consequences of this construction are analysed in Section 5.2.

B.3 BN ARM DAG

The BN arm replaces the layered graph with an explicit Bayesian network sampled to match the empirical structural profile of $\mathcal{G}_{\mathrm{RW}}$. Nodes are introduced in a fixed order $0, \ldots, N-1$; the first n_{c} are roots, and each subsequent node draws its parents only from earlier nodes (guaranteeing acyclicity and making the index order a valid topological order). Two quantities are sampled per node:

- an in-degree k from a fixed empirical distribution concentrated on 1–3 (probabilities $0.42, 0.28, 0.15, 0.08, 0.04, 0.02, 0.01$ for $k = 1, \ldots, 7$, renormalised and capped at a per-graph maximum M matched to the real-world max-parent profile), and

- a set of k parents drawn from a family-specific attachment weighting.

The graph family is drawn per dataset from a categorical distribution whose weights are *themselves* resampled per dataset: logits for four of the five families are drawn as $\mathcal{U}(-3, 5)$ against a reference logit of 1 for the fifth, and softmax-normalised. The mixture is therefore hierarchical rather than a fixed weighting, so individual datasets are frequently dominated by one family while the corpus as a whole spans all five. The families are:

- **Scale-free:** attachment probability $p \propto (\text{outdeg} + 1)^{\alpha}$ with sublinear $\alpha \sim \mathcal{U}(0.3, 0.8)$, concentrating edges on hubs.
- **Uniform / Erdős–R'enyi:** parents chosen uniformly among earlier nodes.
- **Locally-connected:** attachment biased toward recently added nodes, $p \propto e^{-\lambda\,(t-1-\text{index})}$ with $\lambda \sim \mathcal{U}(0.3, 1.5)$, producing long chains.
- **Modular:** nodes are assigned to $\max(2, \lfloor N/12 \rfloor)$ communities and within-community attachment is upweighted $5:1$ relative to cross-community, yielding sparse cross-community links.
- **Forest:** parents drawn from a recent window of 3–12 nodes, producing tree-like structure.

The total node budget is $N = n_{\text{c}} + L \cdot H$, fixed to match the MLP arm and subject to a common computational cap (120 nodes for f30, 180 for f60), so the two arms differ only in *how* the N nodes are wired, not in *how many*.

Because this mixture also makes the BN arm more heterogeneous than the single-mechanism MLP arm, the two differences are entangled in the present ablation (Section 6); we discuss the confound, and the controls that would resolve it, in Appendix D.1.

B.4 Sparsity by construction and the density trap

The constructive scale-free sampler attaches each new node to at most M parents, $k \in \{1, \ldots, M\}$, giving a total edge count linear in N, $|E| = \bar{k}(N-1)$, and hence an average degree centrality

$$\mu_{C_D} \;=\; \frac{1}{N(N-1)} \sum_{v \in V} \deg(v) \;=\; \frac{2\bar{k}(N-1)}{N(N-1)} \;=\; \frac{2\bar{k}}{N}, \tag{11}$$

which *decreases* with N, so target sparsity is guaranteed *by construction*. This is what makes the BN generator tractable at scale, because the naive alternative—sampling dense Erdős–R'enyi graphs and rejecting those that fail a sparsity criterion—is computationally intractable: the fraction of $G(N, p)$ graphs meeting a real-world sparsity target vanishes as N grows, so acceptance rates collapse for $N > 30$. More importantly for our argument, the same linear edge budget bounds the Markov blanket: with in-degree capped at M and expected out-degree $\bar{k} = O(1)$, Eq. (3) gives $|\text{MB}| = O(1)$ independent of N, which is precisely the property the MLP prior lacks (Figure 1a). The preferential-attachment weighting further concentrates connectivity on a small hub set, reproducing the elevated betweenness centrality that distinguishes $\mathcal{G}_{\text{RW}}$, while the per-node in-degree cap reproduces the low max-parent profile. A validation pass—re-measuring parent counts, blanket width, separator cardinality, and betweenness on samples from the BN generator—confirms the generated graphs fall within the $\mathcal{G}_{\text{RW}}$ range on all significant metrics.

B.5 Evaluation-suite augmentation

The high-dimensional regime is sparse in TALENT (only $\sim$30 datasets with ≥ 50 features, 11 with ≥ 100), so we augmented the suite under a rule fixed before observing any results. Starting from the curated OpenML-CC18 suite and a metadata sweep of active OpenML datasets, we retained every classification dataset that (i) was not already present in TALENT, deduplicated both by normalised name and by dataset *family* (stripping AutoML resample suffixes such as `_seed{n}_nrows{n}_...` so that resampled folds of a base dataset count as one family); (ii) had between 50 and 1000 total features (the upper bound excludes widths at which the base model degrades to near-chance); (iii) had at least 2 classes; (iv) was not a flattened-image / pixel dataset (e.g. MNIST-derived) nor a synthetic near-duplicate family; and (v) loaded and split without error. We further capped each near-duplicate family at two datasets to prevent pseudo-replication,

which would violate the independence assumption of the paired significance tests. This produced 52 additional datasets spanning medical, spectroscopy, sensor/activity, molecular, text, and physics domains, of which one had ≤ 50 effective features after cleaning, 21 had 51–100, 23 had 101–300, and 7 had > 300. Each dataset was converted to the base model's on-disk format with an $80/10/10$ stratified train/validation/test split (the validation split is folded into the in-context set at evaluation, matching the base protocol) and mean-imputation of missing numerical values. The augmented suite comprises 219 evaluated datasets after the standard exclusion of the 15 TabPFNv2 development sets. All four checkpoints (both arms, both experiments) were evaluated on the identical augmented suite; the `source` of each dataset (TALENT vs. external) is recorded so that the TALENT-only and external-only subsets in Figure 2(b) can be computed from a single evaluation run. We stress that this rule was fixed internally in advance but not lodged as a public timestamped protocol, so it does not constitute formal pre-registration.

B.6 Training curves

Figure 4 shows the f60 training curves for both arms, each measured on its own prior.

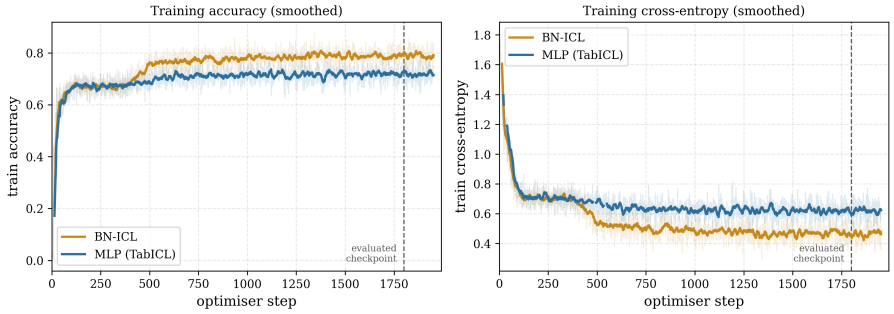


Figure 4: Training accuracy (left) and cross-entropy (right) for the BN and MLP priors in the **f60** experiment, each measured on its own synthetic prior (smoothed; raw values faint). The dashed line marks the evaluated step-1800 checkpoint; training ran to step 1954 (Table 3). Both curves have flattened well before that point, confirming that the comparison is not an undertraining artifact for either arm.

B.7 Hyperparameters

Table 3 lists the full training and architecture configuration. We adopt TabICL's Qu et al. (2025) architecture and training recipe as our starting point and modify it only to fit the reduced-scale, single-variable ablation under a fixed compute budget. All listed values are shared by the MLP and BN arms within each experiment, so no hyperparameter is confounded with the topological variable; the two arms differ solely in `prior_type`.

C The Real-World Graph Corpus $\mathcal{G}_{\text{RW}}$

The real-world corpus used throughout Section 5 and Figure 1 comprises 33 graphs of three provenances, listed in Table 4.

Tier 1: expert-elicited networks (16). Benchmark Bayesian networks from the `bnlearn` repository Scutari (2010), spanning medical diagnosis, engineering, agriculture, meteorology, and genetics; these encode structure elicited from domain experts rather than inferred from data. Two of them, MUNIN and MUNIN1, are the full network and sub-network 1 of the same neuromuscular-diagnosis model; we include both as they differ substantially in size and connectivity, but exclude MUNIN2–4 so that no single family over-represents the corpus.

Tier 2: author-learned reference networks (5). Networks whose structure was learned from real data by their original authors—gene expression Opgen-Rhein & Strimmer (2007), plant-genetics trials Scutari et al. (2014); Scutari (2016), environmental and health monitoring Vitolo et al. (2018),

Table 3: Training and architecture hyperparameters, shared across both arms within each experiment. Only the entries in the *Data / prior* block differ between f30 and f60; all optimisation and architecture settings are identical across both experiments.

Hyperparameter	f30	f60
Data / prior		
Synthetic datasets (train)	10^6	10^6
Epochs	1	1
Min / max features	5 / 30	5 / 60
Max classes	5	5
Node-budget cap	120	180
SCM mechanism activations	8 cheap (App. B.1)	full library (56)
Min / max sequence length	256 / 1024	256 / 1024
Log-uniform sequence length	yes	yes
Sequence length per group	yes	yes
Min / max train fraction	0.2 / 0.9	0.2 / 0.9
Replay small datasets	yes	yes
Optimisation		
Optimiser	AdamW (weight decay $= 0$)	
Learning rate	1×10^{-4}	
LR schedule	cosine with warmup	
Warmup proportion	0.2	
Batch size (global)	512	
Batch size per group	4	
Micro-batch size	1	
Gradient clipping	1.0	
Precision	float32 + AMP	
Total steps	1954	
Evaluation checkpoint	step 1800	
Hardware	$8\times$ GPU, DDP	
Seeds (numpy / torch)	42 / 42	
Architecture		
Embedding dim	128	
Column attention (blocks / heads / inducing pts)	3 / 4 / 128	
Row attention (blocks / heads / CLS tokens)	3 / 8 / 4	
Row-attention RoPE base	10^5	
ICL layers (blocks / heads)	12 / 4	
Feed-forward factor	2	
Norm-first	yes	
Dropout	0.0	
Activation	GELU	
Column feature grouping	disabled	

and viticulture Magrini et al. (2017)—and distributed as reference networks. Several are Gaussian or conditional-Gaussian; we use topology only. These networks also serve as an external check on our implementation: the `bnlearn` repository publishes an average Markov-blanket size and a maximum in-degree for each, and our independently computed values reproduce all ten published figures exactly (e.g. MEHRA 13.75 and 13; ARTH150 3.35 and 6; MAGIC-IRRI 9.97 and 11), confirming that the moral-graph construction of Eq. (5) and our parent-count extraction agree with a reference implementation.

Tier 3: networks we learned (12). Structure and parameters learned from public tabular datasets with `pgmpy` Ankan & Textor (2024): Hill-Climbing search with the AIC-discrete score and Maximum-Likelihood parameter estimation, run without an iteration cap. Continuous variables were discretised beforehand, since the discrete learner requires categorical inputs: numeric columns with at most 20 distinct values were kept at native granularity, and the remainder were binned into five *quantile* intervals—falling back to a mode-versus-rest split for zero-inflated columns where quantile binning degenerates to a single bin. Categorical columns with more than 12 levels had rare levels pooled into an OTHER category, so that a single edge does not become prohibitively expensive under a penalised score; missing values were encoded as an explicit level rather than dropped, so no

rows are lost; and identifier columns (record, family, and subject keys) were removed. All variables are retained as graph nodes, including any left isolated by the search, so the node counts in Table 4 equal the variable counts.

The three tiers are subject to different biases, which is precisely why we use all of them. Hill-Climbing with a penalised score itself favours sparse, bounded-parent graphs, so Tier 3 networks should be read as data-grounded but algorithm-mediated; Tier 1 networks are not subject to that bias, but do reflect human preferences for legible diagrams; Tier 2 networks are learned, but by independent authors using their own methods and domain data. That all three tiers exhibit the same profile—narrow blankets, bounded parents, small separators, size-invariant width—is stronger evidence than any single tier. The corpus also deliberately includes dense counter-examples: MEHRA (13 maximum parents, mean blanket 13.8) and SANGIOVESE (9 maximum parents, mean blanket 11.2) are far denser than typical benchmark networks, and including them *reduces* every measured gap.

Scope. A limitation of $\mathcal{G}_{\mathrm{RW}}$ is that it consists entirely of publicly available networks and datasets. Expert-built benchmark networks and public tabular data may skew toward structures that are simpler, or differently organised, than those found in large proprietary industrial systems; we therefore treat the corpus as a broad but not exhaustive sample of real causal structure.

None of the datasets underlying the Tier 2 and Tier 3 networks overlaps with any dataset in our evaluation suite, from either TALENT or the OpenML augmentation (Appendix B.5); the structural corpus and the predictive-evaluation corpus are therefore disjoint.[1]

D Extended Discussion

This appendix expands three limitations summarised in Section 8 and reports the feature-truncation control referenced there.

D.1 Topology versus generator heterogeneity

We consider the topological reading of our results the more likely explanation, for two reasons. First, the pattern of results tracks the structural predictions rather than diversity as such: the advantage grows monotonically with feature count (Figure 2(a)), which is what the blanket-scaling argument of Section 5.2 predicts, and it is *negative* in the mid-dimensional bucket, which a generic diversity benefit would not produce. Second, a heterogeneity effect would be expected to help roughly uniformly across regimes, since a broader generator distribution should improve coverage of real-world structure everywhere rather than only at high dimensionality. Neither argument is decisive, however, and we cannot exclude a contribution from heterogeneity. Two clean resolutions exist: a single-family BN arm (for instance scale-free only), which removes the heterogeneity difference while retaining the topological one; or a mixture-of-MLPs arm, which adds heterogeneity to the baseline while retaining its topology. Either isolates the remaining variable, and we identify them as the priority follow-up experiments. We note that a single-family control is not a complete decomposition on its own: moving from one family to five changes both the diversity of the generator *and* the set of topologies it can express, so a residual difference between a single-family arm and the mixture would itself require further families to be tested individually before it could be attributed to diversity rather than to the particular structures those families contribute.

D.2 Effect size and scale

Both arms are trained at deliberately reduced scale to isolate the topological variable under a fixed compute budget, and both trail the published TabICL checkpoint in absolute accuracy. We

[1] This disjointness is a good-faith empirical check; given the number of datasets involved, we cannot rule out an undetected duplicate under a different name. We stress, however, that the integrity of our predictive results does not depend on it. The corpus $\mathcal{G}_{\mathrm{RW}}$ enters the pipeline solely through the *aggregate* structural statistics that parameterise the BN generator's sampling families (Appendix B.3); no specific graph structure, conditional-probability table, or dataset is ever used to train or evaluate a model. Any residual overlap could therefore transmit only a few label-free summary scalars (e.g. the median max-parent count), which cannot inflate held-out accuracy. The reported effects are thus invariant to the exact membership of $\mathcal{G}_{\mathrm{RW}}$.

Table 4: Composition of the real-world causal-graph corpus $\mathcal{G}_{\mathrm{RW}}$ (33 graphs; median 33 nodes, range 4–1041). "Nodes" and "Edges" are *measured* from the graph files. Tier 3 networks were learned by us with `pgmpy` Ankan & Textor (2024) (Hill-Climbing search, AIC-d score, ML parameter estimation; see text).

Network	Provenance	Nodes	Edges	Domain / description
Tier 1: expert-elicited benchmark networks (16)				
CHILD	bnlearn Scutari (2010)	20	25	Congenital heart-disease diagnosis
INSURANCE	bnlearn Scutari (2010)	27	52	Car-insurance risk
WATER	bnlearn Scutari (2010)	32	66	Wastewater treatment
MILDEW	bnlearn Scutari (2010)	35	46	Crop mildew management
ALARM	bnlearn Scutari (2010)	37	46	Anaesthesia patient monitoring
BARLEY	bnlearn Scutari (2010)	48	84	Barley crop yield
HAILFINDER	bnlearn Scutari (2010)	56	66	Weather forecasting
HEPAR2	bnlearn Scutari (2010)	70	123	Liver-disorder diagnosis
WIN95PTS	bnlearn Scutari (2010)	76	112	Printer-fault troubleshooting
PATHFINDER	bnlearn Heckerman et al. (1992)	109	195	Lymph-node pathology diagnosis
MUNIN1	bnlearn Scutari (2010)	186	273	Neuromuscular diagnosis (sub-network 1)
ANDES	bnlearn Scutari (2010)	220	338	Intelligent tutoring (physics)
DIABETES	bnlearn Scutari (2010)	413	602	Blood-glucose / insulin management
PIGS	bnlearn Scutari (2010)	441	592	Livestock genetics
LINK	bnlearn Scutari (2010)	714	1125	Genetic linkage analysis
MUNIN	bnlearn Scutari (2010)	1041	1397	Neuromuscular diagnosis (full)
Tier 2: learned from real data by their original authors (5)				
SANGIOVESE	Magrini et al. (2017)	15	55	Viticulture (conditional Gaussian)
MEHRA	Vitolo et al. (2018)	24	71	Air pollution, climate and health
MAGIC-NIAB	Scutari et al. (2014)	44	66	Plant genetics (NIAB wheat)
MAGIC-IRRI	Scutari (2016)	64	102	Plant genetics (IRRI rice)
ARTH150	Opgen-Rhein & Strimmer (2007)	107	150	*Arabidopsis* gene expression
Tier 3: learned by us from public tabular datasets (12)				
Galton Stature	Galton (1886)	4	3	Parent–child height
College Plans	Sewell & Shah (1968)	5	9	College attendance intentions
Auto MPG	Quinlan (1993)	8	7	Automobile fuel economy
PIMA Diabetes	National Institute of Diabetes (1990)	9	11	Diabetes onset prediction
Sachs	Sachs et al. (2005)	11	24	Protein-signalling network
Pittsburgh Bridges	Reich & Fenves (1989)	12	11	Bridge design attributes
Seoul Bike Sharing	Sathishkumar & Cho (2020)	13	31	Bike-rental demand
Adult Census	Becker & Kohavi (1996)	15	30	Income prediction
Credit Approval	Quinlan (1987)	16	25	Credit-card application approval
Hitters	James et al. (2013)	20	22	Baseball salary
South German Credit	sou (2020)	21	34	Consumer credit risk
Student Performance	Cortez (2008)	33	50	School academic outcomes

conjecture—but do not demonstrate—that the high-dimensional deltas we report are informative about what a full-scale, high-feature-count run would show, since our feature caps (30 and 60) sit at or below the regime where the effect is largest. We are deliberate in not asserting this as a lower bound. A much larger model trained on far more data may partially learn to discount the MLP prior's wide conditionals—in effect learning to ignore the parts of the prior that do not generalise—which would narrow rather than widen the gap. The direction is an empirical question that a scaling study, not an argument, would settle.

Our contribution is accordingly methodological: an isolation of DAG topology as a causal factor in prior quality, together with reproducible evidence that real-world-matched topology helps in a predictable, structurally-identifiable regime. A full-scale system trained on the order of the tens of millions of synthetic datasets used by production tabular foundation models, together with a head-to-head comparison against current state-of-the-art tabular methods such as TabICLv2 Qu et al. (2026) and TabPFN-3 **?**, is the natural next step. It is deliberately out of scope here for two reasons: it would not isolate the topological variable, since a full-scale run necessarily changes model capacity, data budget, and prior simultaneously; and a fair competitive evaluation—with hyperparameter search

for the baselines, as the benchmarks themselves perform—would substantially exceed the present format.

D.3 Safeguards on the evaluation-suite augmentation

Because we enriched the evaluation suite specifically in the high-dimensional regime, the aggregate accuracy delta could in principle be biased by the composition of the added datasets. We mitigate this in three ways.

(i) The rule was fixed before observing results. The selection criteria in Appendix B.5 were specified in advance of running any evaluation on the added datasets, and they are stated in terms of dataset metadata (feature count, class count, family, load success) rather than performance. No dataset was added or removed after its delta was known.

(ii) We report the subsets separately. Table 1 gives the merged suite, and Figure 2(b) gives the TALENT-only and external-only slices, so a reader can verify that the effect is not carried by either source alone.

(iii) The central claim rests on the bucketed result, not the aggregate. The headline is the > 100-feature comparison within each experiment, which is a within-regime contrast and therefore insensitive to how many datasets of other widths the suite contains. Its replication on the 32 external datasets alone ($+1.39\%$, CI $[+0.58\%, +2.35\%]$, Holm-significant) is the key safeguard: those datasets are disjoint from TALENT and were selected by the fixed rule, so the effect cannot be an artifact of the base benchmark's composition.

We note that the scarcity of high-dimensional tabular benchmarks is recognised independently: TabBench-Bio **?** was introduced specifically to serve high-dimensional biomedical tables, the regime our augmentation targets for the same reason. Evaluating on it is natural future work.

D.4 The feature-truncation control

Table 5 reports the control described in Section 8. Each of the 43 datasets with more than 100 features was re-evaluated at its full width and at 60 and 30 features under two selection rules; RANDOM is averaged over three seeds per dataset before pairing. The FULL row is a pipeline self-check and reproduces the main-text > 100-feature deltas exactly ($+1.63$pp under f30, $+1.50$ under f60), confirming that the truncated copies are constructed and matched correctly.

Three readings follow. *First*, the effect is not an artifact of exceeding the training feature cap: at 30 TOP-k features the f60 delta remains $+0.86$pp with a bootstrap CI excluding zero and a significant Wilcoxon test. *Second*, at matched width the delta is about twice as large under TOP-k as under RANDOM, in both experiments, indicating that what the BN prior exploits is dependence structure rather than feature count. *Third*, the delta nonetheless decreases monotonically with truncation, so a component of the advantage is genuinely specific to wide tables.

Two caveats. TOP-k selection preserves informative structure *and* discards noise features; our design cannot separate these, so the supported claim is that the advantage survives truncation when the retained features are informative and not when they are arbitrary, rather than that structure preservation alone is responsible. Separately, several datasets in this slice have small test splits (the smallest has seven query rows), so individual deltas are coarsely quantised; this is consistent with the concentrated, non-median-significant pattern reported throughout, and makes the median significance at 30 TOP-k features the more notable. We also note that RANDOM_{30} is slightly *negative* in both experiments (-0.09 and -0.33pp), an unplanned replication of the mid-dimensional negative of Section 7: a 30-feature table with its dependence structure disrupted is precisely the dense, mid-width regime where the MLP prior holds a small edge.

Table 5: Feature-truncation control on the 43 datasets with > 100 features. Δ is BN$-$MLP in percentage points, with a paired bootstrap 95% CI (10^4 resamples) and a two-sided Wilcoxon signed-rank p. FULL is the untruncated self-check and reproduces the main-text deltas exactly; TOP-k retains the highest-mutual-information features, RANDOM an arbitrary subset averaged over three seeds per dataset. Widths 60 and 30 lie inside both training feature caps, so no extrapolation remains.

Exp.	Condition	Width	MLP acc	BN acc	Δ (pp)	CI / Wilcoxon
f30	FULL	>100	0.8148	0.8311	$+1.63$	$[+0.65, +2.78]$, $p = 0.002$
	TOP-k	60	0.8187	0.8255	$+0.68$	$[-0.12, +1.78]$, n.s.
	TOP-k	30	0.8064	0.8103	$+0.39$	$[-0.23, +1.09]$, n.s.
	RANDOM	60	0.7947	0.7984	$+0.37$	$[-0.41, +1.20]$, n.s.
	RANDOM	30	0.7764	0.7755	-0.09	$[-0.72, +0.59]$, n.s.
f60	FULL	>100	0.8170	0.8320	$+1.50$	$[+0.70, +2.34]$, $p = 0.002$
	TOP-k	60	0.8101	0.8225	$+1.24$	$[+0.02, +2.57]$, n.s.
	TOP-k	30	0.7958	0.8044	$\mathbf{+0.86}$	$[\mathbf{+0.34}, \mathbf{+1.45}]$, $\mathbf{p = 0.005}$
	RANDOM	60	0.7815	0.7875	$+0.60$	$[-0.23, +1.62]$, n.s.
	RANDOM	30	0.7614	0.7581	-0.33	$[-1.11, +0.44]$, n.s.