The Missing Structure: When Graph Representations Outperform Tabular Models

Tamara Cucumides
University of Antwerp
tamara.cucumidesfaundez@uantwerp.be

Floris Geerts
University of Antwerp
floris.geerts@uantwerp.be

Abstract

Row-local tabular models excel when labels depend only on per-row attributes. Yet many real-life labels depend on *other rows* (shared values, references, group effects). We ask when explicit cross-row structure becomes *necessary*. Starting from a single table, we construct controlled row-level tasks that require existence or counting of values, and compare (i) strong row-local learners, (ii) the same learners with one-hop neighbor feature aggregation (NFA), and (iii) message passing on graphs induced directly from the table. In this controlled setting, NFA yields small and inconsistent gains over row-local baselines, suggesting that static neighbor summaries are insufficient to recover relational dependencies. Message passing reliably captures the required cross-row logic. These findings reveal a structural difference between tabular and graph learning and suggest that dynamic propagation, rather than static aggregation, is key when targets depend on other rows.

1 Introduction

Most practical machine learning operates on tables: rows as examples, columns as attributes. For years, the strongest tabular systems have been *row-local* pipelines: gradient-boosted trees (GBDTs) and modern multi-layer perceptrons (MLPs) that process examples independently. Under carefully curated i.i.d. evaluation, these models remain hard to beat: When labels are per-row functions, careful validation and ensembling matter more than architectural novelty [McElfresh et al., 2023, Erickson et al., 2025]. Beyond purely row-based approaches, transformer-style tabular foundation models (e.g., TabPFN [Hollmann et al., 2025], TabICL [Qu et al., 2025]) incorporate cross-row mechanisms through attention and in-context learning. Yet, they still do not explicitly follow typed relations or compose multi-hop evidence, capabilities that real-world tables often demand. Indeed, real-world tables often *violate* row independence. Identifiers recur across rows; attributes are shared; and correct predictions depend on whether a value is unique, a key matches a reference, or how many rows fall into the same group. Such dependencies distribute signals across rows, creating a latent relational structure that strictly row-local learners cannot exploit, since they never condition on other rows.

This motivates a focused question: **When does a tabular problem actually require cross-row structure?** We study this under a minimal setup: Holding a single table fixed, we *switch on* cross-row dependencies only through the labels (existence, uniqueness, counting) and compare three levels of structural access: (i) standard row-local baselines, (ii) the same baselines augmented with precomputed one-hop neighbour feature aggregation (NFA) and (iii) shallow message passing on a graph induced by the table (i.e., where rows become nodes and value sharing defines edges). Additionally, the transformer-based model TabPFN serves as a non–row-local reference using inter-row attention but without explicit graph edges.

This design lets us experimentally isolate what kinds of cross-row reasoning each architecture can express. In our controlled tasks, NFA provides limited and inconsistent gains, whereas message passing captures the required relational logic and closes the gap entirely. Rather than a weakness, this delin-

AI for Tabular Data workshop at EurIPS 2025.

eates the boundary: Precomputed one-hop aggregates can help in relational data where features interact, but explicit message propagation becomes essential when labels depend directly on other rows.

2 Background and Related Work

Strong row-local baselines on i.i.d. tables. On curated i.i.d. benchmarks, GBDTs and MLPs remain the strongest tabular learners, with most leaderboard gains attributable to validation and ensembling rather than architectural novelty [Grinsztajn et al., 2022, McElfresh et al., 2023, Salinas and Erickson, 2024]. Erickson et al. [2025] explicitly curate i.i.d. tasks and again find GBDTs and MLPs neck-and-neck, confirming the dominance of row-local methods when examples are independent. These results set a reference point: when labels are truly *per-row*, row-local models are hard to beat.

From tables to graphs. A complementary line of work shows how to turn tables into graphs, either from a single table (with nodes as rows and links based on similarity or value-sharing) or from a full relational database via schema edges [Fey et al., 2024, Li et al., 2025]. On top of these graph constructions, GNNs are applied to enable message passing between related rows. Unlike row-local models, such architectures can naturally model cross-row dependencies when the graph is constructed to reflect shared values or relational structure. The chosen construction *matters*: Cucumides and Geerts [2025] show that representing predictive attributes as nodes can change downstream performance. Still, a formal justification of *what is gained* by turning a table into a graph remains limited; our study takes a first empirical step in that direction.

Tabular methods for graph benchmarks. Recently, the graph learning community has turned its attention towards *relational data*, which is natively tabular. Benchmarks such as RelBench [Robinson et al., 2024] and GraphLand [Bazhenov et al., 2025] build graphs from relational databases, where rows or entities become nodes and schema links define edges. Notably, Bazhenov et al. [2025] include *tabular baselines* enhanced with one-hop neighborhood feature aggregation (NFA) to compare fairly against graph models. Their results show that such NFA baselines can match graph methods on shallow dependencies, while multi-hop or typed relations still favor message passing. We take the complementary view: starting from a *table*, we progressively add graph structure to test when message passing becomes necessary beyond one-hop added features.

Expressivity. The expressiveness of message-passing GNNs is well understood through their connection to color refinement (and 1-WL) and fragments of first-order logic with counting [Morris et al., 2019, Barceló et al., 2020]. This theoretical view clarifies *how* structure expands what can be represented: one-hop aggregation captures local statistics such as degrees or existence, while deeper propagation enables multi-hop and compositional reasoning.

Building on these insights, we compare row-local, NFA-augmented, and message-passing models on a table and its graph representation. Our goal is to provide a setting that isolates—conceptually and empirically—*when and why* structural information becomes necessary for tabular prediction.

3 A Conceptual Gap (Minimal Formal Check)

Row-local models use only per-row features; graph-based models condition on neighbors. This leads to a distinct *sensitivity to table extensions*: if we add or remove rows, some labels (such as uniqueness or reference existence) should change, but a strictly row-local predictor cannot react.

Let A be a set of column names and Dom a value domain. A table $T \subseteq \mathrm{Dom}^A$ is a finite set of rows $r : A \to \mathrm{Dom}$, where each r assigns a value to every column in A. From T, we derive a $G_T = (V, E)$ with (i) one node per row ($V = T$), and (ii) an undirected, typed edge $\{r, s\} \in E$ of type c whenever $r[c] = s[c]$ for some chosen column $c \in A$. A (binary) predictor is a function F that, given a table T, produces per-row outputs $F_T : T \to \{0, 1\}$.

Definition 1 (Extension invariance)**.** *A predictor F is* extension-invariant *if for all tables T and rows $r \in T$, $F_T(r) = F_{\{r\}}(r)$.*

That is, predictions for a row depend only on its own attributes, and not on the presence or absence of other rows.

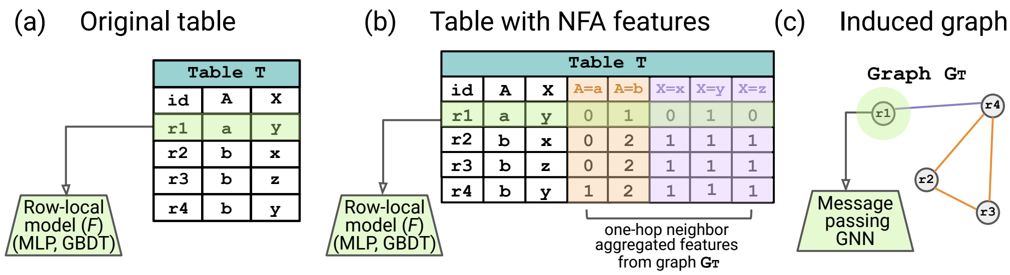

Figure 1: Overview of our three experimental settings: (1) the original table T with a row-local model, (2) T augmented with NFA, and (3) the graph G_T with message passing.

Proposition 1 (Why structure helps)**.** *There exist simple binary labels, such as the* ***uniqueness*** *of a value in a column, that cannot be represented by any extension-invariant predictor in a dataset-independent way, while a single hop of message passing on G_T can compute them exactly.*

Let $f_c(r) = \mathbf{1}\{\,|\{s \in T : s[c] = r[c]\}| = 1\,\}$ denote the uniqueness label for row r in column c. If a duplicate row $s \neq r$ with $s[c] = r[c]$ is added to T, $f(r)$ changes but $F_T(r)$ remains fixed for any extension-invariant F. In the graph G_T, rows sharing c-values are directly connected, so $f_c(r) = \mathbf{1}\{\deg_c(G_T, r) = 0\}$, which is computable by one-hop aggregation over G_T.

This highlights the core gap: row-local predictors are extension-invariant, while message passing is not. Our controlled tasks (*uniqueness, existence, counting*) probe exactly this boundary, tracing a clear hierarchy from *no structure* to *local cues* to *explicit propagation*.

4 Experiments: Structure-Sensitive Tasks, Setup, and Results

We test a simple claim: *does a simple structural channel change what is learnable?*

Setup. We start from a table T (10k rows, 6 columns) with fixed splits and add structure in two ways. (i) *Row-local* tabular baselines operate on the original features only. (ii) *One-hop NFA* augments these baselines with a single round of precomputed neighbor signals. (iii) A *GNN* performs message passing on G_T. All models are fit on the same training split; early stopping and hyperparameter budgets are matched within each family. The configurations are illustrated in Figure 1.

NFA and graph construction. From table T, we derive graph G_T ($10k$ nodes, $\approx 120k$ typed edges) as described in Section 3. For *one-hop neighbor feature aggregation (NFA)*, we precompute simple summaries over each row's neighbors in G_T: for numerical attributes we add mean, min, and max statistics; for categoricals we append per-value neighbor *counts* (color-refinement style). No external data or cross-table links are introduced.

Tasks. We focus on labels that *require cross-row reasoning*, in contrast to i.i.d. benchmarks where row-local methods excel. Each label is defined directly from the table and depends on how many other rows share a pair of given attribute values, making them inherently extension sensitive.

- **Uniqueness within group:** a row is positive if its value in column c_1 appears *exactly once* within rows that share the same value in column c_2,
$$y_c(r) = \mathbf{1}\{\,|\{s \in T : s[c_1, c_2] = r[c_1, c_2]\}| = 1\,\}.$$
- **Counting within group (eq):** positive if its values in c_1 and c_2 occurs *exactly k times*,
$$y_c(r) = \mathbf{1}\{\,|\{s \in T : s[c_1, c_2] = r[c_1, c_2]\}| = k\,\}.$$
- **Counting within group (geq):** positive if its values in c_1 and c_2 occurs *at least k times*,
$$y_c(r) = \mathbf{1}\{\,|\{s \in T : s[c_1, c_2] = r[c_1, c_2]\}| \geq k\,\}.$$

These labels require reasoning over sets of rows (they change when new duplicates or matches are added) so they directly test extension sensitivity. To prevent leakage, all counts used to define $y_c(\cdot)$ are computed *within* each split (train/validation/test) before training or evaluation.

Model	Uniqueness		Counting (eq, $k = 3$)		Counting (geq, $k = 3$)	
	Acc	ROC-AUC	Acc	ROC-AUC	Acc	ROC-AUC
RealMLP (row-local)	0.770	0.540	0.318	0.371	0.374	0.382
LightGBM (row-local)	0.774	0.514	0.506	0.485	0.498	0.500
RealMLP + NFA	0.770	0.451	0.452	0.300	0.380	0.387
LightGBM + NFA	0.778	0.529	0.536	0.452	0.420	0.487
TabPFN	0.786	0.453	0.566	0.474	0.692	0.516
MPGNN (2 L)	**0.811**	**0.738**	**0.806**	**0.771**	**0.998**	**0.998**

Table 1: Accuracy and ROC-AUC on structure-sensitive labels (mean over three seeds). Adding one-hop NFA features yields small and inconsistent gains over row-local baselines in this controlled setting and does not close the gap. A GNN (2 layers) reliably captures the required cross-row counts and achieves the highest scores across tasks.

Models. *RealMLP* [Holzmüller et al., 2024] and *LightGBM* [Ke et al., 2017] are row-local (original features only): both are state-of-the-art on i.i.d. tables [Erickson et al., 2025]. *RealMLP+NFA* and *LightGBM+NFA* receive the same inputs augmented with one-hop signals computed on G_T, collapsing a single aggregation hop into features. A 2-layer heterogeneous SAGE *GNN* [Hamilton et al., 2017] with sum aggregation and ReLU activation runs message passing on G_T. *TabPFN* operates on T without added graph features; it must infer structural patterns from attention alone, whereas the GNN is given the explicit edge structure of G_T.

Results. As expected, row-local models underperform on signals that depend on other rows. Augmenting them with one-hop NFA features produces only marginal and inconsistent changes, often leaving ROC–AUC near chance, because although NFA gives each row its group size, turning these counts into the exact labels requires the model to learn many separate cases across all value groups—something that does not happen in practice. By contrast, a GNN matches or exceeds all alternatives on every task, indicating that a few rounds of propagation suffice to realize the required counting logic at inference time. TabPFN, despite inter-row attention and strong performance on standard i.i.d. benchmarks, remains below the GNN on these structure-sensitive tasks, underscoring that generic cross-row attention is not, by itself, a substitute for targeted message passing on the induced graph. Overall, the pattern is consistent with a structural explanation: in settings where labels hinge on set cardinalities over peers, exposing an actual neighborhood at inference time (via message passing) is necessary; collapsing one-hop aggregates into static features is not enough here.

5 Discussion and Outlook

Causal story. This work takes a first step toward understanding *when* tabular prediction truly requires cross-row structure. We formalized the limitation of row-local models: *extension invariance*, and showed how controlled, structure-sensitive tasks make this gap observable and measurable. One-hop aggregation (NFA) probes this limitation in the narrowest way, adding limited cross-row information but not enough to recover relational logic. Message passing, in contrast, directly relaxes extension invariance and succeeds once dependencies compose beyond a single hop. Together, these results identify structure as a missing inductive bias rather than a side effect of model capacity or tuning.

Practical recipe. Keep standard tabular preprocessing and add a *small* structural layer: build a graph from repeated values or references, try one-hop neighbor features as a simple test, and use one or two layers of message passing if one-hop features are not enough. In many relational datasets, NFA gives cheap gains when structure mixes with other features, but message passing works best when labels depend on other rows. This mirrors results from GraphLand [Bazhenov et al., 2025], where lightweight graph links help, but full propagation is needed for multi-hop reasoning.

Limitations and next steps. Our study is deliberately controlled: a first exploration rather than a comprehensive benchmark. Future work should test real tabular datasets where extension-sensitive dependencies arise (e.g., repeated identifiers, reference joins) and analyze robustness under noisy or spurious edges. On the theoretical side, formalizing the *expressiveness gap* remains open. Clarifying these directions will help establish a broader framework for understanding *when and why* structural reasoning is necessary in tabular learning, and when simple row-local models already suffice.

References

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? In *Advances in Neural Information Processing Systems*, volume 36, pages 76336–76369, 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/file/f06d5ebd4ff40b40dd97e30cee632123-Paper-Datasets_and_Benchmarks.pdf`.

Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. Tabarena: A living benchmark for machine learning on tabular data. *arXiv preprint*, 2025. URL `https://arxiv.org/abs/2506.16791`.

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. URL `https://www.nature.com/articles/s41586-024-08328-6`.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=0VvD1PmNzM`.

Leo Grinsztajn, Edouard Oyallon, and Gael Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, volume 35, pages 507–520, 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf`.

David Salinas and Nick Erickson. Tabrepo: A large scale repository of tabular model evaluations and its automl applications. In Katharina Eggensperger, Roman Garnett, Joaquin Vanschoren, Marius Lindauer, and Jacob R. Gardner, editors, *Proceedings of the Third International Conference on Automated Machine Learning*, volume 256 of *Proceedings of Machine Learning Research*, pages 19/1–30. PMLR, 09–12 Sep 2024. URL `https://proceedings.mlr.press/v256/salinas24a.html`.

Matthias Fey, Weihua Hu, Kexin Huang, Jan Eric Lenssen, Rishabh Ranjan, Joshua Robinson, Rex Ying, Jiaxuan You, and Jure Leskovec. Position: Relational deep learning - graph representation learning on relational databases. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 13592–13607. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/fey24a.html`.

Cheng-Te Li, Yu-Che Tsai, Chih-Yao Chen, and Jay Chiehen Liao. Graph neural networks for tabular data learning: A survey with taxonomy and directions. *ACM Comput. Surv.*, 58(1), September 2025. URL `https://doi.org/10.1145/3744918`.

Tamara Cucumides and Floris Geerts. From features to structure: Task-aware graph construction for relational and tabular learning with GNNs. In *Proceedings of the VLDB 2025 Workshops: Tabular Data Analysis (TaDA)*. VLDB Endowment, 2025. URL `https://www.vldb.org/2025/Workshops/VLDB-Workshops-2025/TaDA/TaDA25_5.pdf`.

Joshua Robinson, Rishabh Ranjan, Weihua Hu, Kexin Huang, Jiaqi Han, Alejandro Dobles, Matthias Fey, Jan E. Lenssen, Yiwen Yuan, Zecheng Zhang, Xinwei He, and Jure Leskovec. Relbench: A benchmark for deep learning on relational databases. In *Advances in Neural Information Processing Systems*, volume 37, pages 21330–21341, 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/25cd345233c65fac1fec0ce61d0f7836-Paper-Datasets_and_Benchmarks_Track.pdf`.

Gleb Bazhenov, Oleg Platonov, and Liudmila Prokhorenkova. Graphland: Evaluating graph machine learning models on diverse industrial data. *arXiv preprint*, 2025. URL `https://arxiv.org/abs/2409.14500`.

Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. Weisfeiler and Leman go neural: Higher-order graph neural networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI)*, pages 4602–4609. AAAI Press, 2019. URL `https://doi.org/10.1609/aaai.v33i01.33014602`.

Pablo Barceló, Egor V. Kostylev, Mikael Monet, Jorge Pérez, Juan Reutter, and Juan Pablo Silva. The logical expressiveness of graph neural networks. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=r1lZ7AEKvB`.

David Holzmüller, Léo Grinsztajn, and Ingo Steinwart. Better by default: Strong pre-tuned mlps and boosted trees on tabular data. In *Advances in Neural Information Processing Systems*, volume 37, pages 26577–26658, 2024. URL `https://proceedings.neurips.cc/paper_files/paper/2024/file/2ee1c87245956e3eaa71aaba5f5753eb-Paper-Conference.pdf`.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf`.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*, volume 31, page 1025–1035, 2017. URL `https://dl.acm.org/doi/pdf/10.5555/3294771.3294869`.