KNOWSTFM: TOWARDS KNOWLEDGE-GROUNDED TABULAR FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Tabular foundation models have advanced deep learning for tabular data by achieving strong performance across many tasks. However, adapting these methods to data-scarce domains is challenging. Many such domains provide curated relational knowledge in the form of knowledge graphs and bases, but how to use this knowledge to improve and steer specialist tabular foundation models remains unclear. We address this problem through **Know**ledge-informed fine-tuning of **T**abular **F**oundation **M**odels (KNOWSTFM), which incorporates relational knowledge directly into feature attention. We evaluate the approach across controlled nanoscale TabPFN checkpoints and the frontier TabPFN-v3 model to study when external structure improves downstream adaptation. During fine-tuning, we use structural priors derived from knowledge graphs to guide feature-to-feature attention, encouraging interactions between features linked in the graph. Our results on 42 biomedical and 13 cross-domain datasets show that this approach yields meaningful gains over standard fine-tuning in domain-specific tasks, while maintaining competitive performance on general-domain tasks.[1]

1 INTRODUCTION

Tabular foundation models (TFMs) have emerged as a leading architecture for solving tabular machine learning tasks (Qu et al., 2025; Hollmann et al., 2025; Ma et al., 2025). These models are typically pretrained on large collections of synthetic tasks designed to mimic real-world datasets, enabling them to develop strong in-context learning capabilities. Instead of training a task-specific model from scratch, a pretrained TFM conditions on labelled support examples and predicts labels for query examples in a single forward pass. This paradigm is especially appealing for small-data regimes, where conventional deep learning models are difficult to train reliably.

The broader foundation model community has shown that adaptation of open-weights models to target domains via fine-tuning can boost the performance of these models across modalities, from vision to large language models (Zhang et al., 2025a). Accessible open-weights models like Gemma (Team et al., 2024), fully open-source initiatives such as OLMo (Groeneveld et al., 2024), and transparent, open-development Pythia-style (Biderman et al., 2023) models have democratised foundation-model research, making it reproducible and inspectable outside large industrial labs. Yet, standard fine-tuning must infer the structure of a new domain primarily from the target data itself, making fine-tuning TFMs in data-scarce domains difficult (Rubachev et al., 2025). This is particularly limiting in scientific domains, where datasets may be small and high-dimensional while substantial relational knowledge already exists outside the observed data. For some domains, like biomedical prediction tasks, tabular features may correspond to genes, drugs, pathways, proteins, or clinical entities, and relations among these entities are available in biomedical knowledge graphs. Such curated knowledge represents years of expert effort. Therefore, our central question is: *Can formal knowledge structure be used to improve and steer tabular foundation models without disrupting their pretrained in-context inference behaviour?*

Strongly pretrained TFMs such as TabPFN-v3 (Grinsztajn et al., 2026) raise an important adaptation problem: while their broad synthetic pretraining provides powerful in-context inference, domain-specific fine-tuning can perturb behaviours that are already useful. External knowledge offers a

[1] Source code is available at `https://anonymous.4open.science/r/kg_tuned_tfms-DBC5`.

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.

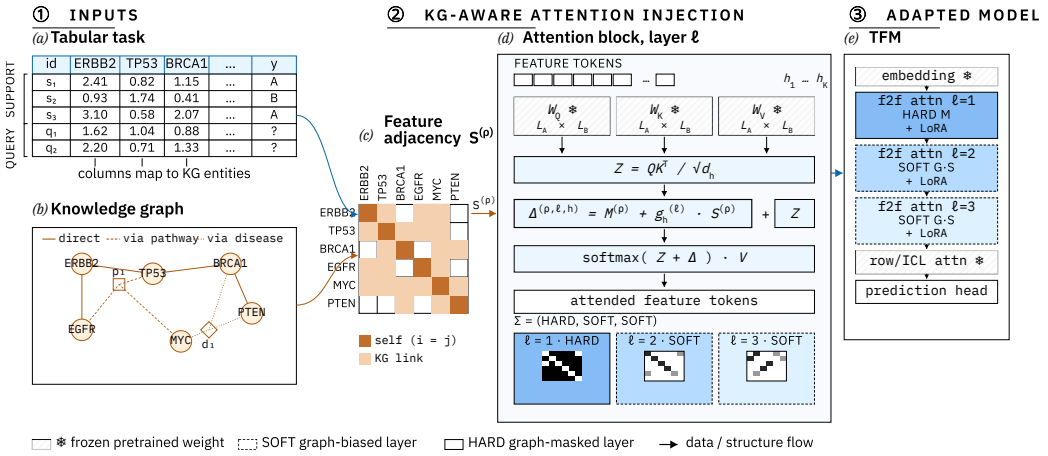


Figure 1: **KG-aware fine-tuning of tabular foundation models.** Given tabular query features (a) and a knowledge graph (b), we derive a feature adjacency matrix from graph relationships (c), and inject this structure into transformer attention through knowledge-informed attention biases and HARD graph-masked / SOFT graph-biased layers (d). The adapted TFM combines frozen pretrained components with lightweight LoRA-based modules to produce downstream predictions (e).

natural way to constrain this adaptation by providing structure that does not need to be inferred from scarce target labels alone. We therefore study knowledge-informed tuning first as a practical mechanism for adapting a frontier TFM, and use nanoscale models to understand the conditions under which it succeeds. Recent nano-TFM models, including nanoTabPFN (Pfefferle et al., 2025) and nanoTabICL (Qu et al., 2025), provide an inspectable testbed in which model capacity, synthetic pretraining prior, and adaptation mechanism can be varied systematically. This combination lets us examine the same knowledge-injection principle across both strongly pretrained frontier models and controlled nanoscale variants.

We study this question through KNOWSTFM, a knowledge-informed adaptation mechanism illustrated in Figure 1. We inject knowledge-graph structure into attention through hard masks or soft additive biases over feature-level interactions, according to the backbone. This framework allows us to examine, in a controlled setting, when and how external knowledge can guide the adaptation of tabular foundation models. We use nanoscale TabPFN variants (Pfefferle et al., 2025) to vary model scale and synthetic prior, and the frontier TabPFN-v3 model to test whether knowledge-informed adaptation extends to a strongly pretrained model. Our overall aim is to understand when structured external knowledge improves domain adaptation and how it can be injected without disrupting pretrained inference behaviour.

Our contributions can therefore be described as:

1. A novel knowledge-informed fine-tuning approach that injects feature-level graph structure directly into a TFM's column attention, using a mix of hard masks and soft gates alongside low-rank adaptation.
2. Empirical evidence that KG-structured attention constraints can improve downstream adaptation across model variants, ranging from small specialist backbones to frontier models.
3. Evidence that LLM-generated graphs offer an alternative to KG grounding when curated coverage is unavailable or incomplete, without requiring explicit entity linking.

2 RELATED WORK

Our work combines several research directions around a specific practical goal: tuning tabular foundation models to use external domain knowledge while preserving pretrained inference behaviour. As such, it relates to prior work in tabular foundation models, domain-knowledge-informed prediction, structural priors, and parameter-efficient fine-tuning, which have been largely studied in isolation.

Tabular foundation models. Recent work has established foundation models as a viable paradigm for tabular deep learning (Hollmann et al., 2023; Qu et al., 2025). TabPFN showed that a transformer pretrained on synthetic tasks can solve small tabular classification problems in a single forward pass, effectively performing in-context learning over training examples (Hollmann et al., 2023). Subsequent work broadened this design space with general tabular transformers and transfer-oriented pretraining, including TransTab and CARTE (Wang & Sun, 2022; Kim et al., 2024), while TabPFN v2 showed that the foundation-model approach can scale to larger and more realistic settings (Hollmann et al., 2025). TabICL extends in-context tabular modelling to larger datasets through a scalable architecture (Qu et al., 2025). On the adaptation side, TuneTables and LoCalPFN show that pretrained tabular models can be specialised efficiently (Feuer et al., 2024; Thomas et al., 2024), and recent fine-tuning studies suggest that downstream adaptation improves performance but can materially alter the pretrained inference mechanism (Rubachev et al., 2025). However, these adaptation methods do not use external knowledge to steer TFM fine-tuning.

Knowledge graphs for tabular and biomedical prediction. Knowledge-informed learning has also been studied in tabular and biomedical settings, although mostly through task-specific architectures rather than adaptation of pretrained tabular transformers. PLATO is close in spirit to our work: it uses an auxiliary knowledge graph to regularise high-dimensional tabular learning by tying feature weights through graph structure (Ruiz et al., 2023) for the drug-to-gene response regression task. In biomedicine, DrugCell encodes Gene Ontology structure directly into a visible neural network for drug response prediction (Kuenzi et al., 2020), while DeepCDR and DRPreter combine graphs, pathway knowledge, and neural architectures for pharmacogenomic prediction (Liu et al., 2020a; Shin et al., 2022). At the data level, resources such as Bioteque and PrimeKG further demonstrate the value of reusable biomedical knowledge graph representations for downstream prediction (Fernández-Torras et al., 2022; Chandak et al., 2023). In contrast, our focus is not on designing a new task-specific architecture from scratch, but on injecting domain knowledge into a pretrained tabular foundation model through attention priors and parameter-efficient updates.

Injecting structure into transformer attention. In NLP, K-BERT introduces a visible matrix that restricts attention according to injected knowledge graph structure, and KnowBERT inserts knowledge modules between transformer layers to recontextualise token representations with entities (Liu et al., 2020b; Peters et al., 2019). GREAT and GraphCodeBERT bias or mask attention using relational graphs over code (Hellendoorn et al., 2020; Guo et al., 2021). These works motivate our use of knowledge-graph-derived structural attention priors, but they do not address pretrained tabular foundation models, where features themselves may correspond to entities in a biomedical or semantic graph. LoRA and adapter-based transfer established that low-rank or bottleneck updates can adapt large pretrained transformers with only a small fraction of trainable parameters (Hu et al., 2022; Houlsby et al., 2019). Building on this idea, KnowLA and KG-Adapter integrate knowledge graph representations into LLM adaptation pipelines, while KBLaM modifies attention to incorporate external knowledge without fully updating the base model (Luo et al., 2024; Tian et al., 2024; Wang et al., 2025). More deeply fused language–graph models such as GreaseLM and DRAGON further show that multi-layer interaction between transformer representations and graph reasoning can be beneficial (Zhang et al., 2022; Yasunaga et al., 2022). As summarised by Pan et al. (2024), these methods form part of a broader trend towards knowledge-graph-enhanced foundation models. Our work extends this trend to the tabular domain.

Preserving in-context learning during fine-tuning. Finally, our use of parameter-efficient updates is motivated by recent evidence that naive fine-tuning can suppress in-context learning behaviour. ProMoT shows that conventional fine-tuning often induces format specialisation that harms generalisation (Wang et al., 2024), while Kotha et al. (2024) argue that downstream tuning may not erase pretrained capabilities outright, but can bias the model away from invoking them. This issue is particularly relevant for TabPFN-style models, whose performance relies on pretrained in-context inference over tabular examples. Our method therefore uses lightweight updates and structured knowledge priors to adapt the model while reducing interference with its pretrained mechanism.

3 KNOWSTFM

Let a tabular task consist of a labelled support set $\mathcal{D}_s = \{(x_i, y_i)\}_{i=1}^{n_s}$ and query inputs $\mathcal{D}_q = \{x_j\}_{j=1}^{n_q}$ (Figure 1a). A pretrained tabular foundation model defines an in-context predictor $p_\theta(y_q \mid x_q, \mathcal{D}_s)$.

We hypothesise that knowledge stored in an external knowledge base (e.g., PrimeKG (Chandak et al., 2023), see Figure 1b) can provide an inductive prior over feature interactions (Figure 1c), regularising adaptation in scarce-label regimes where the support set alone may not identify which columns should interact. We propose KNOWSTFM, a two-component method comprising: (**i**) a graph-retrieval pipeline that maps table columns to a feature-level interaction adjacency (Section 3.1 and Figure 1a-c), and (**ii**) a KG-aware attention injection strategy (Section 3.2 and Figure 1d-e). Section 3.3 details how KNOWSTFM is instantiated to the PFN family.

3.1 TABLE TO KNOWLEDGE GRAPH MAPPING

Let the tabular feature set be $\mathcal{C} = \{c_1, \ldots, c_K\}$. We map columns to KG entities through a partial linking function $\phi : \mathcal{C} \to V \cup \{\varnothing\}$, where $\varnothing$ marks unmapped features, producing S^{ρ}, the adjacency between features. Based on the domain, we propose two mapping mechanisms:

Grounding against expert-curated knowledge. For biomedical datasets, gene-expression features can often be normalised to the HGNC-approved form (`HER2` $\rightarrow$ `ERBB2`), and then linked to a gene-covering knowledge graph like PrimeKG (Chandak et al., 2023). Given the linked nodes, we define a biomedical feature-level adjacency by:

$$
\begin{aligned}
S_{ij}^{(\rho)}=1 \Longleftrightarrow\ & \underbrace{\{\phi(c_i), \phi(c_j)\} \in E_{\mathrm{gg}}}_{\text{direct gene–gene}} \\
& \vee \underbrace{\exists d \in V_{\mathrm{dis}} : \{\phi(c_i), d\} \in E_{\mathrm{gd}}, \{\phi(c_j), d\} \in E_{\mathrm{gd}}}_{\text{shared disease}} \\
& \vee \underbrace{\exists p \in V_{\mathrm{path}} : \{\phi(c_i), p\} \in E_{\mathrm{gp}}, \{\phi(c_j), p\} \in E_{\mathrm{gp}}}_{\text{shared pathway}}
\end{aligned}
$$

where E_{gg}, V_{dis}, and V_{path} denote PrimeKG gene–gene edges, disease nodes, and pathway nodes, respectively. This setting represents an idealised grounding regime: entity names are already close to canonical identifiers, and the main challenge is graph projection rather than entity disambiguation.

General domain grounding. General-domain tabular metadata is often messy or incomplete: curated KGs with native typed entities may be unavailable, and column names are often abbreviated (`trestbps`, `Mg`, `chol`), making careful candidate generation and disambiguation essential. We therefore construct $S^{(\rho)}$ from Wikidata via a four-stage LLM-assisted mapping and disambiguation pipeline, executed once per dataset, that consumes only the dataset name and column descriptions:

1. **Query proposal.** A single Gemini-3 Pro Preview call sees the dataset name and *all* column descriptions, and proposes 3–7 Wikidata search queries per column. Cross-column conditioning is what makes this stage non-trivial: in a glass-composition dataset, `Mg` is proposed as "magnesium" rather than "Madagascar" precisely because the LLM sees the other element columns simultaneously.

2. **Candidate retrieval.** Each proposed query is resolved against the Wikidata REST API; the top candidates are deduplicated and tagged with their unique Wikidata identifiers (QIDs). This step is deterministic and purely API-driven.

3. **LLM disambiguation with abstention.** A second Gemini-3 Pro Preview call sees, per column, the original column description, the dataset context, and the candidate list, and either selects a single QID with a confidence and a rationale, or *abstains*. The LLM is constrained to choose only from the retrieved candidates; it cannot invent a QID that is not in the Wikidata response. Abstention allows the model to leave a feature unmapped rather than guess a QID, as a naive "prompt-the-LLM-for-the-QID" baseline might.

4. **Deterministic edge discovery.** Given the resolved QIDs, edges between feature pairs are obtained by SPARQL queries against the public Wikidata endpoint under three relation policies (direct/1-hop/ancestor;). Every step *after* the LLM disambiguation is deterministic for a fixed Wikidata snapshot; we release the cached `qids` artefact and retrieved graphs to ensure reproducibility despite changes to Wikidata.

3.2 KG-AWARE ATTENTION INJECTION

The structural prior $S^{(\rho)}$ is injected into the pretrained backbone by adding a graph-derived term to the pre-softmax logits of the feature-attention layers, so that attention is biased toward feature pairs connected in the knowledge graph. Crucially, this injection is decoupled from how the backbone is adapted: the same KG term can be used with full fine-tuning or a frozen backbone with lightweight adapters. Throughout, the KG adjacency S^{ρ} and any mask derived from it remain frozen; the KG enters only as an additive term on the feature-attention logits. The full algorithm breakdown can be found in Appendix B.

For a feature-attention layer with queries Q, keys K, values V, and head dimension d_h, we inject a slot-dependent projected graph prior $\Delta^{(\rho)}$ into the standard attention logits $Z = \frac{QK^\top}{\sqrt{d_h}}$ before the softmax:

$$\mathrm{Attn}_{\mathrm{KG}}(Q, K, V) = \mathrm{softmax}\left(Z + \Delta^{(\rho)}\right) V.$$

The form of $\Delta^{(\rho)}$ depends on the layer slot, which we define next.

Per-layer schedule and per-head soft bias. Rather than using a single global injection strength, we assign to each feature-attention block ℓ a slot $\sigma_\ell \in \{\text{HARD}, \text{SOFT}, \text{OFF}\}$, thus yielding $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_L)$. A HARD slot applies a graph mask:

$$M_{ij}^{(\rho)} = \begin{cases} 0, & S_{ij}^{(\rho)} = 1, \\ -C, & S_{ij}^{(\rho)} = 0, \end{cases} \qquad C \gg 0,$$

so that off-graph feature pairs are suppressed in the attention logits. A SOFT slot adds a learnable per-head graph bias:

$$\Delta_{\text{SOFT}}^{(\rho,\ell,h)} = g_h^{(\ell)} S^{(\rho)}, \qquad \text{where} \qquad g_h^{(\ell)} = \sigma(a_h^{(\ell)})\, s_h^{(\ell)}.$$

Here $a_h^{(\ell)}$ is a learned gate parameter and $s_h^{(\ell)}$ is a learned scale parameter. They are initialised as $a_h^{(\ell)} = 0$ and $s_h^{(\ell)} = 1$, so the initial soft bias strength is $g_h^{(\ell)} = 0.5$. An OFF slot leaves the attention logits unchanged. Our main schedule is $\boldsymbol{\sigma} = (\text{HARD}, \text{SOFT}, \ldots, \text{SOFT})$, extended as needed for deeper backbones. Thus, the first adapted block enforces the KG mask, while later blocks use softer KG-guided biases. The KG enters only as an additive term to the pre-softmax attention logits; it is not injected into the value vectors, embedding layer, or feed-forward layers. The graph adjacency $S^{(\rho)}$ and its hard mask remain frozen; backbone weights are frozen under LoRA adaptation. This separates external structure from trainable capacity: the KG determines which feature interactions are encouraged or suppressed, while weight updates provide the task-specific adaptation capacity.

3.3 PFN-FAMILY OF INTERVENTION MODELS

Our main focus is the TabPFN model family, specifically three model variants: small and base NanoTabPFN (under two synthetic-prior regimes), and the frontier TabPFN-v3. Note that in this work we fine-tune all models, but only pretrain the nano variants due to resource limitations. Nevertheless, our pretraining approach is model-scale agnostic. Pretraining is performed following the nanoTabPFN recipe (Pfefferle et al., 2025) (see Appendix A for details). Specifically, each nanoscale checkpoint has parameters $\theta = \theta(b, s, \pi)$ determined by its backbone b, scale s, and synthetic prior regime π. Thus, two models with the same architecture but different pretraining priors are treated as distinct pretrained variants (see Table 5 in Appendix A). The synthetic prior family (see Table 6 in Appendix A) controls the distribution of tabular tasks observed during pretraining. We use two prior data generation regimes:

$$\pi_{\text{small}} : K \in [5, 60], \qquad \pi_{\text{big}} : K \in [40, 200],$$

with a maximum sequence length of 300 and at most 8 classes in both cases. The small and big prior regimes contain 1000 and 2000 generated batches, respectively, both drawn from the same Structural Causal Models (SCM)-based mixed prior family. Thus, they mainly differ in feature-count regime and the amount of generated prior data. By controlling the pretraining of small TFMs, we can test how differently sized synthetic priors generalise to truly unseen downstream tasks. All nanoscale variants in Table 5 in Appendix A use the same optimisation budget. Consequently, differences across matched variants can be attributed to backbone family, model scale, or synthetic prior regime

rather than optimisation budget. This setup lets us examine whether KG-aware adaptation can compensate for weak or mismatched synthetic priors, make better use of support examples, or inject domain structure that scarce labels alone cannot reveal.

4 EXPERIMENTAL SETUP

The experimental analysis is organised around three questions addressing when external knowledge improves adaptation of tabular foundation models: domain adaptability (Q1), generalisability (Q2) and structure recovery (Q3).

Q1: Does domain-matched knowledge-informed tuning improve TFMs? For the main experimental track, we focus on the gene-expression classification of the CUMIDA-42 (Feltes et al., 2019) suite of 42 gene microarray problems. In this setting, feature columns correspond to gene symbols and can be directly matched to PrimeKG (Section 3.1). For the four nanoscale models (see Table 5 in Appendix A) and for the frontier model TabPFN-v3, we compare KG-aware fine-tuning (KNOWSTFM FT) against vanilla full fine-tuning (Vanilla FT), with pretrained zero-shot inference as a baseline. We restrict each dataset to 100 randomly sampled genes (RND100), which lies inside the feature range of the `big` prior and above that of `small` (Table 6). Building on the findings of Rubachev et al. (2025), we consider both full-parameter fine-tuning and low-rank adaptation (LoRA) of a frozen backbone.

Q2: Can general-purpose knowledge transfer to cross-domain tabular tasks? To investigate whether our findings extend beyond biomedicine, we construct a benchmark (xdomain13) of 13 datasets from diverse domains (see Appendix C). In the absence of a curated, domain-specific KG, we apply the LLM-assisted grounding described in Section 3.1 to extract a feature graph from Wikidata. We evaluate the four NanoTabPFN checkpoints and TabPFN-v3 under the same protocol.

Q3: Does the knowledge itself matter? We deconstruct the graph's role by varying its source and content. We first compare the Wikidata-grounded graph on xdomain13 against an LLM-generated feature-to-feature graph, contrasting curated grounding with implicit parametric memory (Petroni et al., 2019). We then replace the knowledge graph with a purely random graph under the same adaptation recipe, testing whether arbitrary structure can reproduce the observed gains. Finally, to establish an oracle reference free from imperfect entity mapping, we utilise Bayesian-network benchmarks (Jiang et al., 2026) (asia, sachs; Figure 3). The known ground-truth causal DAGs in these datasets allow us to benchmark the true underlying graph directly against Wikidata pooling strategies.

Protocol and reporting. For each experiment, we use 3 random seeds and 3 stratified folds, with feature scaling fitted separately within each training fold. We report balanced accuracy as the primary metric, and hyperparameter selection is performed per fold. Both Vanilla FT and KNOWSTFM are fine-tuned with early stopping on a fold-local validation set comprising 20% of the training split, under the same optimisation budget within each backbone family: at most 500 steps with patience 25 for the nanoscale models, and at most 20 epochs with patience 6 through the release trainer for TabPFN-v3. Both use AdamW (Loshchilov & Hutter, 2019) and weight decay 10^{-5}, with the frontier model's KG gates in their own parameter group at learning rate 5×10^{-2}. We optimise the learning rate over the grid $5e-5, 1e-4, 5e-4, 1e-3$, and compare between full-parameter fine-tuning and a fixed Low-Rank Adaptation of the frozen backbone. We apply SOFT biases to the column-aggregator blocks of the frontier model, while employing HARD masks for the small variants. More implementation details are available in Appendix F.

Ablations. We vary three ingredients independently, each under the protocol above. *Graph source*: the curated graph is replaced by an LLM-induced graph or a random graph whose density is itself drawn at random. *Oracle pre-selection*: the 100-column panel is filled with a controlled fraction of random-forest-selected genes, from none (`rnd100`) to all (`top100`), in steps of 25, the oracle being fit on the training split only. *Mechanism and regime*: the two injection mechanisms (hard off-graph mask, soft per-head gate) are crossed with the two adaptation regimes (LoRA, full fine-tuning), with the all-OFF schedule recovering vanilla and graph-free LoRA fine-tuning.

5 RESULTS

Knowledge-informed tuning improves performance on domain-specific tasks (Q1). We report the primary results on CUMIDA-42 in Table 1. First, for the small model variants, our knowledge-

Table 1: Mean balanced accuracy on the CUMIDA-42 gene-expression benchmark with random 100-gene subsets (RND100). **Bold** = best in row. *Pretrained*: zero-shot in-context inference. *Vanilla FT*: full-model fine-tuning. *LoRA FT*: low-rank tuning without graph mask. KNOWSTFM: our proposed method. $^{*}p < 0.05$, $^{**}p < 0.01$ vs Vanilla FT (paired Wilcoxon). ELO: Bradley–Terry rating over per-dataset wins, Vanilla FT anchored at 1000.

Model	Params	Zero-shot	Vanilla FT	LoRA FT	KNOWSTFM
NanoTabPFN small-prior, small	1.13 M	$0.446_{\pm 0.109}$	$0.611_{\pm 0.126}$	$0.599_{\pm 0.122}$	$\mathbf{0.642}^{*}_{\pm 0.157}$
NanoTabPFN small-prior, base	3.72 M	$0.441_{\pm 0.110}$	$0.630_{\pm 0.147}$	$0.610_{\pm 0.132}$	$\mathbf{0.632}_{\pm 0.142}$
NanoTabPFN big-prior, small	1.13 M	$0.442_{\pm 0.117}$	$0.605_{\pm 0.132}$	$0.592_{\pm 0.134}$	$\mathbf{0.629}_{\pm 0.149}$
NanoTabPFN big-prior, base	3.72 M	$0.431_{\pm 0.122}$	$0.621_{\pm 0.145}$	$0.595_{\pm 0.132}$	$\mathbf{0.626}_{\pm 0.146}$
mean over 4 nano ckpts	—	0.440	0.617	0.599	*0.632*
ELO (Vanilla FT = 1000)	—	615	1000	941	**1044**
TabPFN-v3	53 M	$0.862_{\pm 0.126}$	$0.856_{\pm 0.131}$	$0.860_{\pm 0.125}$	$\mathbf{0.863}_{\pm 0.123}^{**}$
ELO	—	1057	1000	1034	**1095**

informed fine-tuning strategy can lead to substantial downstream performance improvements over their zero-shot performance, in some cases up to 20 percentage points. The improvements are also evident relative to standard fine-tuning (Vanilla FT), where KNOWSTFM improves mean balanced accuracy for all four NanoTabPFN checkpoints on average, and the improvement is statistically significant for the small-prior, small checkpoint.

KNOWSTFM also performs better across all four nanoscale variants, compared with LoRA fine-tuning, suggesting that low-rank adaptation alone is not responsible for these gains. This is exactly the setting where KnowsTFM should excel: CUMIDA columns are gene symbols, the feature-level graph is derived from a curated and biologically meaningful KG, and the support sets are small relative to the feature dimension, making it difficult to infer reliable feature interactions from labels alone. Compared with a purely random graph (Figure 2), the curated graph performs better across all five CUMIDA-42 checkpoints, while the cross-domain comparison is more mixed but still favours the curated graph on TabPFN-v3 (Q3).

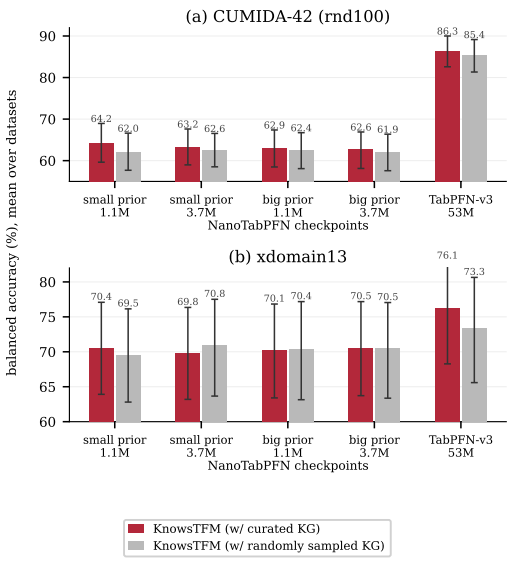


Figure 2: KNOWSTFM with curated domain-specific KG leads to improvements over randomly sampled graphs, with identical workflows. Mean balanced accuracy over (a) biomedical and (b) cross-domain datasets.

Smaller NanoTabPFN models benefit more from the KG injection than the base models (Table 1), and the gain is not amplified by the larger pretraining prior ($+3.1$ vs. $+2.4$ for the small backbone). We interpret this as a backbone-dependent capacity effect: the less structure a model has acquired in pretraining, the more an external sparsity prior may contribute. The choice of mechanism follows a similar pattern, as NanoTabPFNs benefit most clearly from the hard mask, whereas the frontier model benefits from the soft gate and is harmed by the hard mask. Full-parameter fine-tuning with the injected prior does not yield a significant improvement over Vanilla FT at any scale, and extending the soft gate to the frontier model's ICL blocks adds no further gain.

The frontier model, TabPFN-v3, is where full-parameter adaptation is most fragile, and where the structured prior helps preserve pretrained performance (Table 1). Vanilla full-parameter fine-tuning degrades TabPFN-v3 on these small datasets, falling below zero-shot inference. We interpret this as consistent with a scale-and-data mismatch effect: TabPFN-v3 already encodes broad tabular structure through large-scale synthetic-prior pretraining, and fine-tuning it on a single small target dataset may

disturb this robust in-context behaviour rather than improve it. This mirrors observations in large language models, where fine-tuning on narrow or newly introduced knowledge can degrade the use of pretrained knowledge (Gekhman et al., 2024; Li et al., 2024).

Knowledge-informed tuning recovers this loss: KNOWSTFM improves over Vanilla FT ($p < 0.01$) and exceeds zero-shot performance, albeit slightly. LoRA alone recovers part of the degradation but remains below zero-shot performance. This pattern is consistent with the injected structure acting as an anchor on feature attention, limiting how far adaptation moves from the pretrained solution while still allowing task-specific updates. These results suggest that *structured adaptation* may be preferable to full fine-tuning for a strongly pretrained model, with the same pattern also appearing on the cross-domain tasks (Q2). We performed an exploratory dataset-level analysis (Table 20 in Appendix I), and did not find any particular dataset property that remains associated with the gain after multiple-testing correction. Note that the smallest variants show descriptive positive correlations with sample size and samples per class.

Table 2: Mean balanced accuracy on CUMIDA-42 across feature-selection regimes, averaged over 42 datasets. Gains are significant under partial feature selection. Rank: mean per-dataset rank (1 = best). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ vs. Vanilla FT (paired Wilcoxon).

Regime	ZS	vFT	KNOWSTFM
`rnd100`	0.440 ±0.111	0.617 ±0.134	**0.632** ±0.145
(all random)	rank 2.81	rank 1.69	rank **1.50**
`top25rnd75`	0.440 ±0.109	0.670 ±0.140	**0.689** ±0.134 *
	rank 2.88	rank 1.64	rank **1.48**
`top50rnd50`	0.439 ±0.108	0.698 ±0.134	**0.710** ±0.131 *
	rank 2.98	rank 1.69	rank **1.33**
`top100`	0.440 ±0.107	0.729 ±0.140	**0.732** ±0.128
(fully selected)	rank 2.98	rank **1.50**	rank 1.52

To test whether the gain depends on how features are chosen, we additionally vary the feature-selection regime (Table 2, further results in Appendix D): starting from our main setting of 100 randomly sampled genes (`rnd100`), we progressively replace random genes with selected top-relevant ones (`top25rnd75`, `top50rnd50`, `top100`). As expected, replacing random genes with relevance-selected ones improves absolute performance, with KNOWSTFM providing additional gains. Consistent with our reasoning, the benefit of knowledge grounding is greatest for the smallest variants (1.1M) and decreases with increasing model capacity. Note that we use `rnd100` for the KG-vs-random comparison in Q3 to avoid an additional relevance-selection prior.

General-purpose knowledge transfers, but only at the frontier (Q2). Evaluating the same protocol on xdomain13 (Table 3) reveals a strong dependence on the backbone. On the nanoscale models, KG-grounded adaptation differs from vanilla fine-tuning by between -0.8 and +0.5 percentage points. Because a paired test across 13 datasets can only reliably resolve differences of 1.4 to 2.8 points, these small effects fall below the benchmark's detection limit and remain statistically unresolved. On the frontier model, however, the effect is clearer and constitutes the strongest cross-domain evidence for knowledge-informed adaptation in our study. Specifically, we find that while vanilla full fine-tuning degrades TabPFN-v3 performance (relative to its zero-shot inference), KNOWSTFM leads to a more stable behaviour and recovers this loss ($p < 0.01$). Our results show no substantial

Table 3: Cross-domain transfer on the thirteen heterogeneous datasets (xdomain13). Balanced accuracy, mean over datasets. KNOWSTFM variants use the per-model recipe fixed on CUMIDA: hard masking for nano variants and soft masking for the frontier; LLM replaces the Wikidata-grounded graph by the LLM-induced graph under the same recipe. Stars: paired Wilcoxon vs. vanilla FT across datasets ($^{*}p < .05$, $^{**}p < .01$). ELO: Bradley–Terry rating over per-dataset wins with Vanilla FT anchored at 1000. **Bold**: best adapted method in the row.

Model	Params	Zero-shot	Vanilla FT	KNOWSTFM	KNOWSTFM (LLM)
NanoTabPFN small-prior, small	1.13 M	$0.408_{\pm.144}$	$0.700_{\pm.125}$	$\mathbf{0.704}_{\pm.122}$	$0.703_{\pm.120}$
NanoTabPFN small-prior, base	3.72 M	$0.408_{\pm.144}$	$0.706_{\pm.129}$	$0.698_{\pm.120}$	$\mathbf{0.712}_{\pm.127}$
NanoTabPFN big-prior, small	1.13 M	$0.406_{\pm.148}$	$0.696_{\pm.127}$	$\mathbf{0.701}_{\pm.124}$	$0.700_{\pm.125}$
NanoTabPFN big-prior, base	3.72 M	$0.408_{\pm.145}$	$0.710_{\pm.131}$	$0.705_{\pm.121}$	$\mathbf{0.712}_{\pm.117}$
mean (four nano checkpoints)		0.408	0.703	0.702	0.707
ELO (Vanilla FT = 1000)		248	**1000**	938	958
TabPFN-v3	53 M	$0.758_{\pm.133}$	$0.746_{\pm.134}$	$\mathbf{0.761}^{**}_{\pm.132}$	$0.761^{*}_{\pm.133}$
ELO		1100	1000	**1250**	**1250**

performance differences between Wikidata-grounded and LLM-generated graphs. Thus, under the tested adaptation recipe, the cross-domain benefit is concentrated in TabPFN-v3. Analysing the Wikidata mapping, we find that grounding coverage ranges from 8% to 100%, and none of the 13 datasets produces a direct feature-to-feature Wikidata edge. The resulting graphs therefore rely on pooled neighbours and vary widely in density (Table 21, Appendix J). This variability makes the grounding pipeline itself a source of uncertainty and motivates Q3, where we test whether the observed gains depend on the source and exact structure of the injected prior.

LLMs can act as the knowledge base (Q3). We next examine how these gains depend on the source of the structural prior, given that general-domain grounding can be fragile. To this end, we evaluate whether an LLM's parametric memory can serve directly as the knowledge base by generating the feature-to-feature graph entirely via prompting. This LLM-generated graph performs comparably to the Wikidata approach across backbones (Table 3) and occasionally exceeds it, all while requiring no explicit entity linking and exhibiting a lower, more stable mean density of 14.7%. These results indicate that knowledge-informed tuning can be effectively introduced to uncovered domains simply by prompting an LLM for the feature adjacency.

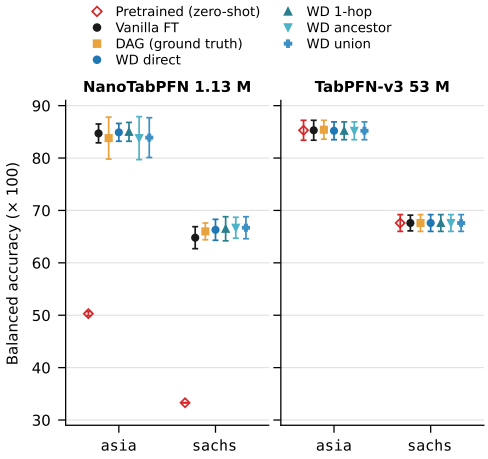


Figure 3: Comparison of structural priors on synthetic Bayesian-network benchmarks. Ground-truth DAGs and Wikidata-derived graphs yield similar performance.

Finally, to test whether exact graph structure provides an advantage, we evaluate synthetic Bayesian-network benchmarks, where the ground-truth causal graph is known and feature-to-entity mapping is exact (Figure 3). At each model size, KNOWSTFM performs similarly across structural priors. Because these synthetic benchmarks are drawn from SCMs similar to the pretraining priors (Zhang et al., 2025b), the lack of separation may partly reflect structure already captured during pretraining. Together with our previous findings, these results suggest that, while structured attention priors can be beneficial, its benefits do not depend strongly on exact semantic edges. When curated KG coverage is limited, LLM-generated graphs therefore can provide a practical alternative for constructing the external structure.

Limitations. KNOWSTFM requires tabular data populated with metadata, which may be unavailable for some tasks. Mapping to knowledge graphs (KGs) can be an ambiguous task, and this ambiguity may propagate to the tuning stage (Ye & Mitchell, 2025). Next, the benefits of KG-informed adaptation depend heavily on the synthetic prior used during pretraining; if the pretraining prior already captures the relevant dependency structures, additional (external) KG priors may not yield further improvements. Additionally, TFM architectures do not all expose feature-level attention in the same manner (see Appendix F). While some models provide a natural column-attention interface, others primarily attend over rows or support-query examples. Consequently, the optimal KG injection mechanism may differ across backbones. Finally, the selected training dynamic setup (number of steps, optimiser choice) might be suboptimal across different model families and sizes, potentially affecting performance, especially for the large models (Rubachev et al., 2025).

6 CONCLUSION

We presented KNOWSTFM, a knowledge-informed framework for adapting tabular foundation models through graph-guided feature attention. Across CUMIDA-42, the gains depend strongly on the checkpoint and adaptation mechanism, with the clearest improvements on the smallest NanoTabPFN models and on TabPFN-v3. On heterogeneous cross-domain tasks, gains are concentrated in TabPFN-v3, while LLM-generated feature graphs provide a practical alternative to Wikidata grounding. Random and oracle graph controls further show that the additional benefit of specific semantic edges remains unresolved. Together, these results position structured attention priors as a promising, but backbone- and regime-dependent, mechanism for adapting pretrained tabular models.

AI USE STATEMENT

We leverage generative AI as a component of the preprocessing pipeline and to assist in preparing the manuscript, so we separate the two uses.

As part of the research artefact. Two elements of KNOWSTFM call a large language model at experiment time. The agentic mapper of Section 3.1 uses an LLM to propose candidate entity identifiers for each column name before those candidates are resolved against Wikidata, and the LLM-generated graph evaluated in Table 3 is produced by prompting an LLM directly for feature-to-feature relations without any grounding step. Both use `gemini-3-pro-preview`. These calls are cached and released with the code, and their outputs are reported as measured quantities rather than accepted as correct: Table 21 gives per-dataset coverage, edge counts and densities for both graph sources.

As an aid to the authors. We used an AI assistant to review parts of the code and analyse it for bugs, to prepare HPC job scripts that make efficient use of our GPU allocation, to improve the design of figures and tables, and to help edit the text. We have reviewed all AI-assisted work, and every number reported in the paper is regenerated from the raw per-fold outputs by scripts included in the supplementary material. We take responsibility for the final content of this paper, including text, claims and artefacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available benchmark data and introduces no new data collection. The CUMIDA-42 gene-expression panels are de-identified public microarray datasets curated for benchmarking; the thirteen cross-domain tables are standard UCI and OpenML datasets; the Bayesian-network benchmarks are synthetic samples from published networks. No human subjects were recruited and no personally identifying information was accessed.

REPRODUCIBILITY STATEMENT

The experimental protocol is detailed in Section 4, while the exact hyperparameters, optimiser settings, step budgets, and per-backbone schedules are provided in Appendix F.1. Algorithm 1 outlines the complete procedure, parameterised so that every arm in the ablation grid is a specific instance of it. Data splits and feature panels are drawn once, frozen in a manifest, and shared identically across all methods and backbones, ensuring no experimental arm differs by anything other than the intervention under test. Each arm undergoes the same four-point learning-rate grid search, tuned per fold on a validation split excluded from fitting. The best-performing configuration is then refit on the full training split before evaluation on the held-out test set; evaluating three seeds across three stratified folds yields nine runs per dataset.

Per-dataset results for every checkpoint, feature regime, and arm are provided in Appendix D and Appendix H rather than being reported solely in aggregate, ensuring any reported mean can be recomputed and any individual result inspected. Appendix L applies the identical protocol to a second backbone family. The graph construction pipeline, including the entity-mapping step, cached knowledge graphs, and LLM outputs, is released so that feature graphs do not need to be regenerated to reproduce the reported numbers. Finally, anonymised source code, frozen manifests, cached graphs, and raw per-fold result files are provided as supplementary material.

REFERENCES

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10:67, 2023. doi: 10.1038/s41597-023-01960-3.

Bruno César Feltes, Eduardo Bassani Chandelier, Bruno Iochins Grisci, and Márcio Dorn. Cumida: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. *Journal of Computational Biology*, 26(4):376–386, 2019.

Adrià Fernández-Torras, Miquel Duran-Frigola, Martino Bertoni, Martina Locatelli, and Patrick Aloy. Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the bioteque. *Nature communications*, 13(1):5304, 2022.

Benjamin Feuer, Robin Tibor Schirrmeister, Valeriia Cherepanova, Chinmay Hegde, Frank Hutter, Micah Goldblum, Niv Cohen, and Colin White. Tunetables: Context optimization for scalable prior-data fitted networks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FOfU3qhcIG.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7765–7784, 2024.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Mihir Manium, Shi Bin Hoo, Magnus Bühler, Anurag Garg, et al. Tabpfn-3: Technical report. *arXiv preprint arXiv:2605.13986*, 2026.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. OLMo: Accelerating the science of language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15789–15809, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.841. URL https://aclanthology.org/2024.acl-long.841/.

Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie Liu, Long Zhou, Nan Duan, Alexey Svyatkovskiy, Shengyu Fu, Michele Tufano, Shao Kun Deng, Colin Clement, Dawn Drain, Neel Sundaresan, Jian Yin, Daxin Jiang, and Ming Zhou. GraphCodeBERT: Pre-training code representations with data flow. In *International Conference on Learning Representations*, 2021.

Vincent J. Hellendoorn, Charles Sutton, Rishabh Singh, Petros Maniatis, and David Bieber. Global relational models of source code. In *International Conference on Learning Representations*, 2020.

Steffen Herbold. Autorank: A python package for automated ranking of classifiers. *Journal of Open Source Software*, 5(48):2173, 2020. doi: 10.21105/joss.02173. URL https://doi.org/10.21105/joss.02173.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations*, 2023. Originally presented at the Table Representation Learning Workshop at NeurIPS 2022.

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*, 637:319–326, 2025. doi: 10.1038/s41586-024-08328-6. URL https://doi.org/10.1038/s41586-024-08328-6.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Xiangjian Jiang, Nikola Simidjievski, and Mateja Jamnik. Tabstruct: Measuring structural fidelity of tabular data. In *The Fourteenth International Conference on Learning Representations*, 2026.

Myung Jun Kim, Leo Grinsztajn, and Gael Varoquaux. CARTE: Pretraining and transfer for tabular learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23843–23866. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/kim24d.html.

Suhas Kotha, Jacob Springer, and Aditi Raghunathan. Understanding catastrophic forgetting in language models via implicit inference. In *International Conference on Learning Representations*, 2024.

Brent M. Kuenzi, Jisoo Park, Samson H. Fong, Kyle S. Sanchez, John Lee, Jason F. Kreisberg, Jianzhu Ma, and Trey Ideker. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell*, 38(5):672–684, 2020. doi: 10.1016/j.ccell.2020.09.014.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. Revisiting catastrophic forgetting in large language model tuning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4297–4308, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.249. URL https://aclanthology.org/2024.findings-emnlp.249/.

Qiao Liu, Zhiqiang Hu, Rui Jiang, and Mu Zhou. DeepCDR: A hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*, 36(Supplement_2):i911–i918, 2020a. doi: 10.1093/bioinformatics/btaa822.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-BERT: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 2901–2908, 2020b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Xindi Luo, Zequn Sun, Jing Zhao, Zhe Zhao, and Wei Hu. KnowLA: Enhancing parameter-efficient finetuning with knowledgeable adaptation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 7153–7166, 2024.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Jesse Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L Caterini, and Maks Volkovs. Tabdpt: Scaling tabular foundation models on real data. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 172692–172722. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/fc0e3f908a2116ba529ad0a1530a3675-Paper-Conference.pdf.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024. doi: 10.1109/TKDE.2024.3352100.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 43–54, 2019.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2463–2473, 2019.

Alexander Pfefferle, Johannes Hog, Lennart Purucker, and Frank Hutter. nanotabpfn: A lightweight and educational reimplementation of tabpfn. *arXiv preprint arXiv:2511.03634*, 2025.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data, 2025.

Ivan Rubachev, Akim Kotelnikov, and Nikolay Kartashev. On finetuning tabular foundation models, 2025. URL https://arxiv.org/abs/2506.08982.

Camilo Ruiz, Hongyu Ren, Kexin Huang, and Jure Leskovec. High dimensional, tabular deep learning with an auxiliary knowledge graph. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 26348–26371. Curran Associates, Inc., 2023. doi: 10.52202/075280-1146. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/53dd219b6b11abc8ce523921c18c7a3e-Paper-Conference.pdf.

Jihye Shin, Yinhua Piao, Dongmin Bang, Sun Kim, and Kyuri Jo. Drpreter: Interpretable anticancer drug response prediction using knowledge-guided graph neural networks and transformer. *International Journal of Molecular Sciences*, 23(22), 2022. ISSN 1422-0067. doi: 10.3390/ijms232213919. URL https://www.mdpi.com/1422-0067/23/22/13919.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Valentin Thomas, Junwei Ma, Rasa Hosseinzadeh, Keyvan Golestan, Guangwei Yu, Maksims Volkovs, and Anthony L. Caterini. Retrieval & fine-tuning for in-context tabular models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=337dHOexCM.

Shiyu Tian, Yangyang Luo, Tianze Xu, Caixia Yuan, Huixing Jiang, Chen Wei, and Xiaojie Wang. KG-Adapter: Enabling knowledge graph integration in large language models through parameter-efficient fine-tuning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3813–3828, 2024.

Xi Wang, Taketomo Isazawa, Liana Mikaelyan, and James Hensman. Kblam: Knowledge base augmented language model. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 51629–51658, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/803485352e61e3ebf41221e4776c9fd4-Paper-Conference.pdf.

Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit Dhillon, and Sanjiv Kumar. Two-stage llm fine-tuning with less specialization and more generalization. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 20380–20398, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/597254dc45be8c166d3ccf0ba2d56325-Paper-Conference.pdf.

Zifeng Wang and Jimeng Sun. TransTab: Learning transferable tabular transformers across tables. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D. Manning, Percy Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Christophe Ye and Cassie S. Mitchell. LLM as entity disambiguator for biomedical entity-linking. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 301–312, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-252-7. doi: 10.18653/v1/2025.acl-short.25. URL https://aclanthology.org/2025.acl-short.25/.

Dan Zhang, Tao Feng, Lilong Xue, Yuandong Wang, Yuxiao Dong, and Jie Tang. Parameter-efficient fine-tuning for foundation models. *arXiv preprint arXiv:2501.13787*, 2025a.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and Jure Leskovec. GreaseLM: Graph reasoning enhanced language models. In *International Conference on Learning Representations*, 2022.

Xiyuan Zhang, Danielle Maddix Robinson, Junming Yin, Nick Erickson, Abdul Fatir Ansari, Boran Han, Shuai Zhang, Leman Akoglu, Christos Faloutsos, Michael Mahoney, Tony Hu, Huzefa Rangwala, George Karypis, and Yuyang (Bernie) Wang. Mitra: Mixed synthetic priors for enhancing tabular foundation models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 15795–15840. Curran Associates, Inc., 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/177d68f4adef163b7b123b5c5adb3c60-Paper-Conference.pdf.

A PER-CHECKPOINT PRETRAINING CONFIGURATION

All 8 nanoscale variants of Table 5 are pretrained with an identical optimisation budget (see Table 4). Table 6 details synthetic prior families used for nanoscale pretraining.

Table 4: Pretraining configuration per nano variant. Steps $\times$ Batch $=$ Tasks seen $= 12{,}000$ synthetic tabular tasks consumed during pretraining. Prior pool is the number of distinct synthetic tasks in the HDF5 prior dump. Passes = Tasks seen / Prior pool, i.e. how many times each prior task is seen on average (the loader wraps the pointer when it reaches the end of the file). Wallclock is per checkpoint on a single GPU. AdamW, learning rate 10^{-4}, fixed seed 2402.

Backbone	Prior	Steps	Batch	Tasks seen	Prior pool	Passes	Wall (s)
NanoTabICL	`tabicl_small`	3,000	4	12,000	1,000	12	464–665
NanoTabICL	`tabicl_big`	3,000	4	12,000	2,000	6	464–670
NanoTabPFN	`tabicl_small`	3,000	4	12,000	1,000	12	442–648
NanoTabPFN	`tabicl_big`	3,000	4	12,000	2,000	6	446–654

Table 5: Nanoscale pretrained model variants. Each architecture-scale pair is pretrained under both synthetic prior families, separating backbone, scale, and prior-family effects.

Backbone	Prior family	Scale	Embedding	Blocks / layers	Parameters
NanoTabICL	`tabicl_small`	small	48	col/row/ICL: 2/2/2; heads: 4/4/4	848,770
		base	80	col/row/ICL: 3/3/3; heads: 4/4/4	3,302,742
	`tabicl_big`	small	48	col/row/ICL: 2/2/2; heads: 4/4/4	848,770
		base	80	col/row/ICL: 3/3/3; heads: 4/4/4	3,302,742
NanoTabPFN	`tabicl_small`	small	128	layers: 4; heads: 4; MLP: 512	1,129,994
		base	192	layers: 6; heads: 6; MLP: 768	3,717,514
	`tabicl_big`	small	128	layers: 4; heads: 4; MLP: 512	1,129,994
		base	192	layers: 6; heads: 6; MLP: 768	3,717,514

Table 6: Synthetic prior families used for nanoscale pretraining. Both priors come from the same SCM-based mixed generator family, but differ in feature range and amount of generated prior data.

Prior family	Size	Batches	Feature range	Max length	Max classes
`tabicl_small`	148 MB	1000	5–60	300	8
`tabicl_big`	1.1 GB	2000	40–200	300	8

B KNOWSTFM ALGORITHM

In this appendix, we outline the algorithm used for knowledge-informed adaptation of the TFMs.

Algorithm 1 Knowledge-informed adaptation of a tabular foundation model. A frozen structural prior S is injected into feature-attention logits according to a per-block schedule, while the backbone is adapted using either LoRA or full fine-tuning.

Require: pretrained TFM f_θ with L feature-attention blocks
Require: training data $\mathcal{D}_{\text{tr}}$, feature names $c_{1:K}$, structural-prior source G and source configuration ξ
Require: schedule $\boldsymbol{\sigma} \in \{\text{HARD}, \text{SOFT}, \text{OFF}\}^L$, regime $R \in \{\text{LORA}, \text{FULL}\}$
Require: LoRA rank r, scale α, learning rates $\boldsymbol{\eta}$, budget T, mask constant $C = 10^4$

1. Construct structural prior

$S \leftarrow \text{BUILDSTRUCTURALPRIOR}(c_{1:K}, G, \xi)$ ▷ source-specific construction; Section 3.1
$S_{ii} \leftarrow 1 \; \forall i$ ▷ retain self-connections

2. Configure KG injection

$\Phi \leftarrow \emptyset$
for $\ell = 1, \ldots, L$ **do**
  if $\sigma_\ell = \text{OFF}$ **then**
    $M^{(\ell,h)} \leftarrow 0 \quad \forall h$
  else if $\sigma_\ell = \text{HARD}$ **then**
    $M_{ij}^{(\ell,h)} \leftarrow \begin{cases} 0, & S_{ij} = 1, \\ -C, & S_{ij} = 0, \end{cases} \quad \forall h$
  else
    $a^{(\ell)} \leftarrow \mathbf{0}, \; s^{(\ell)} \leftarrow \mathbf{1} \in \mathbb{R}^H; \quad \Phi \leftarrow \Phi \cup \{a^{(\ell)}, s^{(\ell)}\}$
    $M_{ij}^{(\ell,h)} \equiv \text{sigmoid}(a_h^{(\ell)}) s_h^{(\ell)} S_{ij}$ ▷ recomputed each forward pass; $g_h^{(\ell)} = 0.5$ initially
  end if
end for

3. Configure adaptation

if $R = \text{LORA}$ **then**
  for each adapted W in Q, K, V, out_proj, and FFN **do**
    initialise L_A, L_B with $L_B = 0; \quad W_{\text{eff}} \leftarrow W + \frac{\alpha}{r} L_B L_A$
    $\Phi \leftarrow \Phi \cup \{L_A, L_B\}$
  end for
else
  $\Phi \leftarrow \Phi \cup \theta$ ▷ full-parameter fine-tuning
end if

4. Episodic fine-tuning

for $t = 1, \ldots, T$ **do**
  sample context/query sets $(\mathcal{S}_c, \mathcal{S}_q)$ from $\mathcal{D}_{\text{tr}}$
  $Z^{(\ell,h)} \leftarrow Q^{(\ell,h)} K^{(\ell,h)\top} / \sqrt{d_h} + M^{(\ell,h)}$ for all adapted feature-attention blocks
  $\hat{y}_q \leftarrow f_\theta(X_{\mathcal{S}_c}, y_{\mathcal{S}_c}, X_{\mathcal{S}_q})$
  $\mathcal{L} \leftarrow \text{CE}(\hat{y}_q, y_{\mathcal{S}_q})$
  $\Phi \leftarrow \text{ADAMWSTEP}(\Phi, \nabla_\Phi \mathcal{L}, \boldsymbol{\eta})$
end for
return adapted f_θ

For NanoTabPFN, the selected LoRA recipe uses (HARD, SOFT, SOFT, SOFT) for the four-block 1.13M checkpoints and (HARD, SOFT, SOFT, SOFT, SOFT, SOFT) for the six-block 3.72M checkpoints. For TabPFN-v3, we use an all-SOFT schedule on the three adapted column-aggregator blocks, with a separate learning rate for the gate parameters. Setting $R = \text{FULL}$ gives the corresponding full-fine-tuning variants, while an all-OFF schedule recovers the graph-free baseline. Non-feature slots, including the target column and the register columns of TabPFN-v3, receive neither a HARD mask nor a SOFT bias. Early stopping and validation follow the protocol of Section 4. Table 14 ablates the resulting grid.

C Dataset selection and metadata

This appendix documents the inclusion criteria and resulting metadata for the three dataset families used in Section 4.

C.1 Biomedical: CUMIDA-42

CUMIDA (Feltes et al., 2019) ships 69 cancer microarray datasets in its public catalog. We retain the 42 panels that satisfy all of the following criteria, applied programmatically and audited per dataset: (i) a non-empty symbol-mapped expression matrix exists, (ii) the target defines a non-trivial classification task with at least two classes and enough minority-class samples for stratified 3-fold cross-validation, and (iii) the panel contains at least $N \geq 12$ patient samples. We do not otherwise filter on sample count; the smallest retained panel is `Breast_GSE26910` with $N = 12$, and the largest is `Leukemia_GSE28497` with $N = 281$. Table 7 gives the resulting CUMIDA-42 suite. "N" is the number of patient samples, "K_{full}" is the number of symbol-mapped expression features before fold-local selection, "cls" is the number of outcome classes, and "maj. %" is the majority-class share.

Sample feature names (first ten gene symbols of `Leukemia_GSE9476` after probe-to-symbol normalisation, before fold-local RF feature selection): `TP53`, `BRCA1`, `KMT2A`, `RUNX1`, `FLT3`, `NPM1`, `IDH1`, `IDH2`, `DNMT3A`, `TET2` (illustrative; the actual selected gene set is fold-specific).

Feature pre-processing. CUMIDA panels contain between 6,820 and 20,779 symbol-mapped expression features, which exceeds the synthetic-prior feature range of our nanoscale checkpoints, $K \in [40, 200]$. We therefore apply fold-local feature selection. On each training fold, we fit a random forest classifier (Breiman, 2001) with 100 trees and maximum depth 10, select the top-T features by Gini importance, and add R disjoint random features, yielding $K = T + R$. The headline arm uses $T = 0, R = 100$; Appendix D reports sensitivity sweeps for $\{T = 100, R = 0\}$, $\{T = 50, R = 50\}$, and $\{T = 25, R = 75\}$.

C.2 Cross-domain UCI/OpenML (general-knowledge)

The 13 datasets in Table 8 were selected by applying three constraints: (i) each dataset must provide feature names or descriptions with enough natural-language semantics for the agentic Wikidata mapper of Section 3.1; (ii) each dataset must have $K \leq 50$ raw features, so all columns can enter $S^{(\rho)}$ without fold-local feature selection; and (iii) each dataset must contain at least $N \geq 150$ samples and at least five samples in each minority class. The retained suite spans $8 \leq K \leq 41$, from `hepatitis` with $N = 155$ to `stroke` with $N = 4{,}909$.

We do not filter on Wikidata-mapping coverage before running experiments. Low-coverage datasets such as `QSAR-biodeg` and `dermatology` are retained deliberately, since these failures are part of the empirical setting studied in Table 21 in Appendix J.

Sample feature names (illustrative, from the glass dataset):

`glass`. `RI` (refractive index), `Na` (sodium), `Mg` (magnesium), `Al` (aluminum), `Si` (silicon), `K` (potassium), `Ca` (calcium), `Ba` (barium), `Fe` (iron).

Feature pre-processing. All K columns are used verbatim. We label-encode categorical columns and standardise numeric columns using training-fold statistics only. For datasets with more than 1,000 samples, namely `stroke`, `QSAR-biodeg`, and `steel-plates-fault`, we use a fixed-seed stratified subsample of 1,000 examples, shared across all method variants.

Table 7: The 42 CUMIDA cancer panels used in Table 1. Source: `https://sbcb.inf.ufrgs.br/cumida`.

#	Dataset	Cancer	N	K_{full}	cls (maj. %)
1	Bladder_GSE31189	Bladder	85	20,779	2 (56%)
2	Brain_GSE15824	Brain	37	20,779	4 (32%)
3	Brain_GSE50161	Brain	108	20,779	4 (43%)
4	Breast_GSE10797	Breast	66	12,939	3 (42%)
5	Breast_GSE26910	Breast	12	20,779	2 (50%)
6	Breast_GSE42568	Breast	116	20,779	2 (87%)
7	Breast_GSE45827	Breast	151	20,779	6 (27%)
8	Breast_GSE7904	Breast	45	20,779	3 (47%)
9	Colorectal_GSE21510	Colorectal	147	20,779	3 (71%)
10	Colorectal_GSE32323	Colorectal	33	20,779	2 (52%)
11	Colorectal_GSE41328	Colorectal	18	20,779	2 (56%)
12	Colorectal_GSE44861	Colorectal	105	12,939	2 (50%)
13	Colorectal_GSE77953	Colorectal	55	12,939	4 (31%)
14	Colorectal_GSE8671	Colorectal	63	20,779	2 (51%)
15	Gastric_GSE19826	Gastric	24	20,779	2 (50%)
16	Gastric_GSE79973	Gastric	20	20,779	2 (50%)
17	Leukemia_GSE14317	Leukemia	25	12,939	2 (72%)
18	Leukemia_GSE22529_U133A	Leukemia	52	12,939	2 (79%)
19	Leukemia_GSE22529_U133B	Leukemia	52	10,184	2 (79%)
20	Leukemia_GSE28497	Leukemia	281	12,939	7 (26%)
21	Leukemia_GSE63270	Leukemia	101	20,779	2 (59%)
22	Leukemia_GSE71935	Leukemia	46	20,779	2 (80%)
23	Leukemia_GSE9476	Leukemia	64	12,939	5 (41%)
24	Liver_GSE14520_U133_2	Liver	41	12,939	2 (54%)
25	Liver_GSE22405	Liver	48	12,939	2 (50%)
26	Liver_GSE60502	Liver	36	12,939	2 (50%)
27	Liver_GSE62232	Liver	91	20,779	2 (89%)
28	Lung_GSE18842	Lung	90	20,779	2 (51%)
29	Lung_GSE19804	Lung	114	20,779	2 (51%)
30	Lung_GSE27262	Lung	48	20,779	2 (50%)
31	Lung_GSE7670	Lung	51	12,939	2 (53%)
32	Ovary_GSE6008	Ovary	98	12,939	4 (42%)
33	Pancreatic_GSE16515	Pancreatic	51	20,779	2 (71%)
34	Prostate_GSE26910	Prostate	12	20,779	2 (50%)
35	Prostate_GSE46602	Prostate	49	20,779	2 (71%)
36	Prostate_GSE55945	Prostate	17	20,779	2 (59%)
37	Prostate_GSE6919_U95Av2	Prostate	124	9,013	2 (50%)
38	Prostate_GSE6919_U95B	Prostate	124	6,820	2 (52%)
39	Renal_GSE53757	Renal	143	20,779	2 (50%)
40	Renal_GSE6344_U133A	Renal	20	12,939	2 (50%)
41	Renal_GSE6344_U133B	Renal	20	10,184	2 (50%)
42	Renal_GSE66270	Renal	28	20,779	2 (50%)

Table 8: The 13 cross-domain UCI/OpenML datasets used in Table 3. "N" = total samples, "K" = features, "cls" = classes, "maj. %" = majority-class share. The agentic v2 Wikidata mapper of Section 3.1 produces the $S^{(\rho)}$ adjacency for each dataset; mapping coverage and edge statistics are reported in Table 21 in Appendix J.

Dataset	N	K	cls	maj. %
automobile	159	25	2	73.6
cirrhosis	418	17	3	55.5
dermatology	366	34	6	30.6
diabetes_pima	768	8	2	65.1
glass	214	9	6	35.5
glioma	839	23	2	58.0
heart-h	294	13	2	63.9
heart-statlog	270	13	2	55.6
hepatitis	155	19	2	79.4
QSAR-biodeg	1055	41	2	66.3
SPECT	267	22	2	79.4
steel-plates-fault	1941	27	7	34.7
stroke	4909	10	2	95.7

Table 9: Feature-selection regimes per checkpoint on CUMIDA-42 (balanced accuracy $\times 100$, mean over 42 datasets; each dataset is the mean of 3 seeds $\times$ 3 folds with per-fold leak-free learning-rate selection). Δ: mean paired difference to vanilla fine-tuning over datasets, paired Wilcoxon ($^{*}p < .05$, $^{**}p < .01$). **Bold**: better of vFT and KNOWSTFM in the row.

Checkpoint	Regime	N	ZS	vFT	KNOWSTFM	Δ vs. vFT
NanoTabPFN small-prior, small	`rnd100`	42	44.6	61.1	**64.2**	$+3.11^{*}$
	`top25rnd75`	42	44.7	66.1	**69.8**	$+3.71^{**}$
	`top50rnd50`	42	44.9	69.8	**71.6**	$+1.84$
	`top100`	42	45.1	73.1	**74.9**	$+1.79$
NanoTabPFN small-prior, base	`rnd100`	42	44.1	63.0	**63.2**	$+0.18$
	`top25rnd75`	42	44.3	**68.0**	68.0	-0.01
	`top50rnd50`	42	44.3	**70.5**	69.8	-0.65
	`top100`	42	44.6	**73.5**	71.8	-1.71^{*}
NanoTabPFN big-prior, small	`rnd100`	42	44.2	60.5	**62.9**	$+2.40$
	`top25rnd75`	42	44.0	67.0	**69.9**	$+2.98^{*}$
	`top50rnd50`	42	43.9	69.5	**71.8**	$+2.36^{*}$
	`top100`	42	43.9	71.7	**73.4**	$+1.68$
NanoTabPFN big-prior, base	`rnd100`	42	43.1	62.1	**62.6**	$+0.54$
	`top25rnd75`	42	42.9	66.8	**67.9**	$+1.05$
	`top50rnd50`	42	42.5	69.5	**70.6**	$+1.08$
	`top100`	42	42.7	**73.4**	72.5	-0.83
TabPFN-v3 (53M)	`rnd100`	42	86.2	85.6	**86.3**	$+0.75^{**}$
	`top25rnd75`	42	89.3	89.4	**89.5**	$+0.12$
	`top50rnd50`	42	89.6	89.5	**89.6**	$+0.05$
	`top100`	42	90.2	90.0	**90.2**	$+0.21^{*}$

D CUMIDA-42 FULL RESULTS

This appendix reports every CUMIDA-42 number behind Tables 1 and 2: all 42 datasets $\times$ five checkpoints, for each of the four feature-selection regimes defined in Section 4. KNOWSTFM uses the per-model recipe fixed in Appendix G. Two points that the pooled summary in the main text cannot show are worth stating before the tables.

We find that the gain is not uniform across checkpoints. Table 9 decomposes the pooled result. It is carried by the two 1.1M models, which improve on vanilla fine-tuning in every regime ($+1.7$ to $+3.7$ points), while the two 3.7M models are flat and turn slightly negative once features are fully relevance-selected (-1.7 on the small-prior base model at `top100`, $p < 0.05$). TabPFN-v3 has little headroom in any regime: its zero-shot accuracy already matches or exceeds vanilla fine-tuning throughout. The regime trend in the main text should therefore be read as a property of the smallest backbones rather than of the family as a whole.

Table 10: Full results on CUMIDA-42, regime `rnd100` (main setting): all 42 datasets × five models. Balanced accuracy (×100), mean$_{\pm \text{std}}$ over the nine (seed, fold) cells, learning rate chosen per fold on validation; the *mean* row averages over the 42 datasets. ZS: zero-shot; vFT: vanilla full fine-tuning; KG: KnowsTFM with the per-model recipe (hard mask with LoRA on the four NanoTabPFN checkpoints, soft gate with LoRA on TabPFN-v3). **Bold**: KG > vFT.

Dataset	small-prior, small			small-prior, base			big-prior, small			big-prior, base			TabPFN-v3		
	ZS	vFT	KG	ZS	vFT	KG	ZS	vFT	KG	ZS	vFT	KG	ZS	vFT	KG
Bladder_GSE31189	$48.8_{\pm 4.6}$	$48.6_{\pm 5.0}$	$\mathbf{51.4}_{\pm 6.5}$	$50.0_{\pm 6.1}$	$49.6_{\pm 4.0}$	$46.3_{\pm 5.8}$	$50.0_{\pm 0.0}$	$51.2_{\pm 5.5}$	$48.5_{\pm 6.4}$	$50.0_{\pm 0.0}$	$47.4_{\pm 4.0}$	$\mathbf{48.2}_{\pm 5.7}$	$51.3_{\pm 5.9}$	$50.4_{\pm 6.1}$	$\mathbf{53.3}_{\pm 6.6}$
Brain_GSE15824	$33.8_{\pm 9.7}$	$52.1_{\pm 13.7}$	$50.2_{\pm 9.1}$	$25.9_{\pm 2.6}$	$56.9_{\pm 7.1}$	$48.1_{\pm 10.3}$	$33.8_{\pm 6.1}$	$56.5_{\pm 16.3}$	$46.1_{\pm 13.1}$	$23.1_{\pm 6.4}$	$53.2_{\pm 17.2}$	$45.6_{\pm 9.9}$	$71.5_{\pm 14.4}$	$64.6_{\pm 18.5}$	$\mathbf{71.5}_{\pm 14.4}$
Brain_GSE50161	$25.0_{\pm 0.0}$	$52.2_{\pm 11.4}$	$\mathbf{52.2}_{\pm 10.1}$	$25.0_{\pm 0.0}$	$54.3_{\pm 16.4}$	$\mathbf{64.4}_{\pm 14.6}$	$25.0_{\pm 0.0}$	$42.6_{\pm 11.5}$	$\mathbf{60.9}_{\pm 15.8}$	$25.0_{\pm 0.0}$	$49.0_{\pm 15.2}$	$\mathbf{60.1}_{\pm 11.1}$	$95.1_{\pm 2.2}$	$94.9_{\pm 2.2}$	$\mathbf{95.1}_{\pm 2.2}$
Breast_GSE10797	$33.3_{\pm 0.0}$	$37.6_{\pm 4.8}$	$\mathbf{39.0}_{\pm 6.4}$	$32.6_{\pm 2.1}$	$36.3_{\pm 5.9}$	$\mathbf{41.2}_{\pm 8.4}$	$32.9_{\pm 1.2}$	$40.2_{\pm 5.4}$	$38.6_{\pm 5.5}$	$30.7_{\pm 8.8}$	$40.4_{\pm 7.6}$	$35.5_{\pm 7.7}$	$59.9_{\pm 11.6}$	$58.3_{\pm 9.8}$	$\mathbf{59.9}_{\pm 11.6}$
Breast_GSE26910	$50.0_{\pm 0.0}$	$52.8_{\pm 7.9}$	$\mathbf{55.6}_{\pm 10.4}$	$50.0_{\pm 0.0}$	$47.2_{\pm 7.9}$	$\mathbf{50.0}_{\pm 0.0}$	$61.1_{\pm 12.4}$	$52.8_{\pm 7.9}$	$52.8_{\pm 14.2}$	$50.0_{\pm 0.0}$	$50.0_{\pm 11.8}$	$50.0_{\pm 0.0}$	$80.6_{\pm 19.6}$	$80.6_{\pm 19.6}$	$80.6_{\pm 19.6}$
Breast_GSE42568	$50.0_{\pm 0.0}$	$66.1_{\pm 14.0}$	$\mathbf{84.1}_{\pm 13.5}$	$50.0_{\pm 0.0}$	$85.1_{\pm 11.8}$	$78.0_{\pm 19.3}$	$50.0_{\pm 0.0}$	$74.5_{\pm 13.3}$	$\mathbf{78.5}_{\pm 14.5}$	$50.0_{\pm 0.0}$	$76.4_{\pm 18.1}$	$\mathbf{84.7}_{\pm 10.7}$	$92.8_{\pm 4.4}$	$92.8_{\pm 4.4}$	$92.8_{\pm 4.4}$
Breast_GSE45827	$16.7_{\pm 0.0}$	$62.2_{\pm 7.8}$	$\mathbf{70.7}_{\pm 7.1}$	$16.8_{\pm 0.4}$	$64.1_{\pm 5.7}$	$\mathbf{69.0}_{\pm 8.2}$	$16.5_{\pm 0.4}$	$65.3_{\pm 10.7}$	$\mathbf{68.5}_{\pm 5.8}$	$16.5_{\pm 0.9}$	$59.7_{\pm 13.7}$	$\mathbf{63.7}_{\pm 9.8}$	$90.1_{\pm 2.1}$	$90.8_{\pm 3.2}$	$89.8_{\pm 2.2}$
Breast_GSE7904	$33.3_{\pm 0.0}$	$45.1_{\pm 12.4}$	$\mathbf{54.7}_{\pm 16.7}$	$33.3_{\pm 0.0}$	$48.5_{\pm 10.6}$	$41.5_{\pm 19.2}$	$33.3_{\pm 0.0}$	$44.4_{\pm 11.6}$	$42.6_{\pm 14.0}$	$27.5_{\pm 5.7}$	$40.4_{\pm 16.4}$	$40.2_{\pm 14.9}$	$91.6_{\pm 9.0}$	$91.1_{\pm 8.7}$	$91.1_{\pm 8.7}$
Colorectal_GSE21510	$33.3_{\pm 0.0}$	$79.1_{\pm 8.0}$	$\mathbf{81.2}_{\pm 12.5}$	$33.3_{\pm 0.0}$	$79.0_{\pm 9.6}$	$\mathbf{82.3}_{\pm 10.8}$	$33.3_{\pm 0.0}$	$78.3_{\pm 10.2}$	$\mathbf{81.2}_{\pm 11.5}$	$33.3_{\pm 0.0}$	$78.4_{\pm 10.9}$	$\mathbf{80.5}_{\pm 11.2}$	$99.9_{\pm 0.3}$	$99.9_{\pm 0.3}$	$99.9_{\pm 0.3}$
Colorectal_GSE32323	$50.0_{\pm 0.0}$	$62.2_{\pm 19.0}$	$\mathbf{78.3}_{\pm 17.7}$	$50.0_{\pm 0.0}$	$65.2_{\pm 19.3}$	$\mathbf{81.5}_{\pm 14.0}$	$49.8_{\pm 11.8}$	$66.3_{\pm 13.2}$	$\mathbf{73.7}_{\pm 16.4}$	$53.1_{\pm 13.4}$	$68.0_{\pm 15.0}$	$\mathbf{81.7}_{\pm 12.9}$	$95.9_{\pm 4.6}$	$95.9_{\pm 4.6}$	$95.9_{\pm 4.6}$
Colorectal_GSE41328	$50.0_{\pm 0.0}$	$67.1_{\pm 16.0}$	$55.6_{\pm 15.7}$	$50.0_{\pm 0.0}$	$56.5_{\pm 17.5}$	$\mathbf{57.9}_{\pm 15.3}$	$52.8_{\pm 18.4}$	$60.2_{\pm 19.2}$	$57.9_{\pm 20.5}$	$55.6_{\pm 15.7}$	$61.1_{\pm 19.6}$	$59.3_{\pm 17.8}$	$98.1_{\pm 5.2}$	$98.1_{\pm 5.2}$	$98.1_{\pm 5.2}$
Colorectal_GSE44861	$50.0_{\pm 0.0}$	$61.6_{\pm 7.9}$	$\mathbf{73.7}_{\pm 7.1}$	$50.0_{\pm 0.0}$	$68.8_{\pm 3.0}$	$68.0_{\pm 7.4}$	$51.3_{\pm 3.6}$	$63.0_{\pm 5.9}$	$\mathbf{71.0}_{\pm 9.1}$	$56.2_{\pm 3.8}$	$66.0_{\pm 6.2}$	$\mathbf{70.8}_{\pm 7.2}$	$81.0_{\pm 9.5}$	$81.0_{\pm 9.5}$	$81.0_{\pm 9.5}$
Colorectal_GSE77953	$28.0_{\pm 4.9}$	$51.8_{\pm 18.2}$	$47.1_{\pm 14.7}$	$25.6_{\pm 4.4}$	$56.7_{\pm 9.4}$	$\mathbf{59.4}_{\pm 11.6}$	$24.4_{\pm 7.5}$	$46.6_{\pm 11.4}$	$\mathbf{50.4}_{\pm 11.9}$	$25.0_{\pm 0.0}$	$59.0_{\pm 11.9}$	$54.5_{\pm 12.7}$	$83.6_{\pm 7.7}$	$83.6_{\pm 8.3}$	$83.6_{\pm 7.7}$
Colorectal_GSE8671	$50.0_{\pm 0.0}$	$73.7_{\pm 13.6}$	$\mathbf{84.1}_{\pm 9.8}$	$50.0_{\pm 0.0}$	$90.9_{\pm 4.2}$	$80.8_{\pm 18.9}$	$47.4_{\pm 6.0}$	$82.5_{\pm 11.1}$	$\mathbf{86.9}_{\pm 8.3}$	$46.4_{\pm 6.9}$	$83.3_{\pm 11.9}$	$80.6_{\pm 12.2}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Gastric_GSE19826	$50.0_{\pm 0.0}$	$56.9_{\pm 8.6}$	$\mathbf{62.5}_{\pm 11.8}$	$50.0_{\pm 0.0}$	$55.6_{\pm 8.6}$	$55.6_{\pm 11.9}$	$61.1_{\pm 9.2}$	$58.3_{\pm 11.8}$	$56.9_{\pm 8.6}$	$47.2_{\pm 7.9}$	$54.2_{\pm 5.9}$	$\mathbf{59.7}_{\pm 9.8}$	$69.4_{\pm 10.4}$	$70.8_{\pm 11.8}$	$70.8_{\pm 11.8}$
Gastric_GSE79973	$50.5_{\pm 1.3}$	$64.8_{\pm 12.4}$	$56.9_{\pm 11.1}$	$50.0_{\pm 0.0}$	$70.4_{\pm 17.2}$	$63.9_{\pm 10.2}$	$45.4_{\pm 11.9}$	$64.4_{\pm 16.3}$	$59.7_{\pm 10.8}$	$46.3_{\pm 10.5}$	$72.2_{\pm 12.1}$	$66.2_{\pm 14.5}$	$86.6_{\pm 8.5}$	$86.6_{\pm 8.5}$	$86.6_{\pm 8.5}$
Leukemia_GSE14317	$50.0_{\pm 0.0}$	$60.2_{\pm 13.5}$	$\mathbf{65.7}_{\pm 22.4}$	$50.0_{\pm 0.0}$	$63.0_{\pm 13.7}$	$58.3_{\pm 16.7}$	$50.0_{\pm 0.0}$	$55.6_{\pm 13.6}$	$\mathbf{57.4}_{\pm 14.4}$	$50.0_{\pm 0.0}$	$60.2_{\pm 13.5}$	$51.9_{\pm 5.2}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Leukemia_GSE22529_U133A	$50.0_{\pm 0.0}$	$48.8_{\pm 2.1}$	$\mathbf{53.4}_{\pm 8.3}$	$50.0_{\pm 0.0}$	$53.1_{\pm 8.5}$	$\mathbf{57.4}_{\pm 12.7}$	$50.0_{\pm 0.0}$	$52.9_{\pm 7.2}$	$\mathbf{55.4}_{\pm 8.8}$	$50.0_{\pm 0.0}$	$51.5_{\pm 9.0}$	$\mathbf{56.4}_{\pm 8.6}$	$93.2_{\pm 7.1}$	$93.2_{\pm 7.1}$	$\mathbf{93.7}_{\pm 6.6}$
Leukemia_GSE22529_U133B	$50.0_{\pm 0.0}$	$49.6_{\pm 7.1}$	$\mathbf{51.0}_{\pm 4.3}$	$50.0_{\pm 0.0}$	$54.3_{\pm 5.8}$	$53.8_{\pm 8.7}$	$50.0_{\pm 0.0}$	$49.4_{\pm 4.0}$	$\mathbf{54.1}_{\pm 9.0}$	$50.0_{\pm 0.0}$	$53.6_{\pm 8.5}$	$\mathbf{54.0}_{\pm 8.7}$	$87.2_{\pm 9.7}$	$87.2_{\pm 9.7}$	$\mathbf{88.6}_{\pm 8.1}$
Leukemia_GSE28497	$14.2_{\pm 0.2}$	$39.4_{\pm 6.6}$	$\mathbf{43.6}_{\pm 5.7}$	$14.3_{\pm 0.0}$	$44.6_{\pm 4.9}$	$\mathbf{50.1}_{\pm 4.8}$	$14.3_{\pm 0.0}$	$36.5_{\pm 11.4}$	$\mathbf{40.9}_{\pm 10.8}$	$14.3_{\pm 0.0}$	$39.3_{\pm 9.3}$	$\mathbf{50.2}_{\pm 6.1}$	$76.5_{\pm 3.1}$	$76.7_{\pm 3.0}$	$76.5_{\pm 3.1}$
Leukemia_GSE63270	$50.0_{\pm 0.0}$	$71.1_{\pm 20.6}$	$\mathbf{91.4}_{\pm 6.1}$	$50.0_{\pm 0.0}$	$87.7_{\pm 15.3}$	$\mathbf{89.6}_{\pm 9.9}$	$50.0_{\pm 0.0}$	$65.7_{\pm 17.1}$	$\mathbf{84.5}_{\pm 9.5}$	$50.0_{\pm 0.0}$	$72.7_{\pm 17.5}$	$\mathbf{88.1}_{\pm 15.0}$	$99.7_{\pm 0.8}$	$99.7_{\pm 0.8}$	$99.7_{\pm 0.8}$
Leukemia_GSE71935	$50.0_{\pm 0.0}$	$50.9_{\pm 2.7}$	$47.2_{\pm 9.4}$	$50.0_{\pm 0.0}$	$50.7_{\pm 10.7}$	$\mathbf{51.1}_{\pm 2.6}$	$50.0_{\pm 0.0}$	$52.3_{\pm 6.6}$	$48.8_{\pm 4.0}$	$50.0_{\pm 0.0}$	$46.2_{\pm 8.8}$	$\mathbf{50.5}_{\pm 1.3}$	$76.9_{\pm 16.2}$	$72.3_{\pm 17.4}$	$\mathbf{76.9}_{\pm 16.2}$
Leukemia_GSE9476	$20.0_{\pm 0.0}$	$74.2_{\pm 11.2}$	$67.0_{\pm 9.4}$	$20.0_{\pm 0.0}$	$70.7_{\pm 5.9}$	$\mathbf{77.2}_{\pm 9.3}$	$20.0_{\pm 0.0}$	$61.6_{\pm 15.4}$	$60.2_{\pm 15.7}$	$20.0_{\pm 0.0}$	$62.5_{\pm 11.5}$	$\mathbf{64.3}_{\pm 22.0}$	$98.2_{\pm 1.9}$	$98.2_{\pm 1.9}$	$98.2_{\pm 1.9}$
Liver_GSE14520_U133_2	$50.0_{\pm 0.0}$	$81.1_{\pm 10.7}$	$\mathbf{85.9}_{\pm 10.7}$	$50.0_{\pm 0.0}$	$76.8_{\pm 11.4}$	$\mathbf{79.4}_{\pm 10.0}$	$45.0_{\pm 6.5}$	$81.9_{\pm 8.6}$	$\mathbf{82.8}_{\pm 7.7}$	$49.3_{\pm 2.0}$	$87.5_{\pm 8.7}$	$83.3_{\pm 9.8}$	$98.4_{\pm 3.1}$	$98.4_{\pm 3.1}$	$98.4_{\pm 3.1}$
Liver_GSE22405	$50.0_{\pm 0.0}$	$65.3_{\pm 7.3}$	$60.4_{\pm 7.8}$	$50.0_{\pm 0.0}$	$65.3_{\pm 9.8}$	$64.6_{\pm 7.8}$	$36.8_{\pm 3.5}$	$66.7_{\pm 9.8}$	$64.6_{\pm 10.2}$	$36.1_{\pm 4.9}$	$63.2_{\pm 9.5}$	$\mathbf{66.0}_{\pm 10.7}$	$78.5_{\pm 9.4}$	$77.1_{\pm 8.8}$	$77.1_{\pm 8.8}$
Liver_GSE60502	$50.0_{\pm 0.0}$	$77.8_{\pm 14.2}$	$68.5_{\pm 12.9}$	$50.0_{\pm 0.0}$	$78.7_{\pm 13.1}$	$68.5_{\pm 11.0}$	$34.3_{\pm 10.7}$	$75.9_{\pm 13.9}$	$66.7_{\pm 16.2}$	$36.1_{\pm 11.1}$	$76.9_{\pm 12.3}$	$68.5_{\pm 12.3}$	$96.3_{\pm 5.7}$	$95.4_{\pm 5.7}$	$\mathbf{96.3}_{\pm 5.7}$
Liver_GSE62232	$50.0_{\pm 0.0}$	$55.7_{\pm 11.2}$	$\mathbf{84.2}_{\pm 19.7}$	$50.0_{\pm 0.0}$	$61.2_{\pm 11.9}$	$\mathbf{67.0}_{\pm 19.0}$	$50.0_{\pm 0.0}$	$54.0_{\pm 8.7}$	$\mathbf{85.3}_{\pm 15.8}$	$50.0_{\pm 0.0}$	$60.6_{\pm 15.4}$	$\mathbf{65.0}_{\pm 21.3}$	$99.2_{\pm 1.8}$	$99.2_{\pm 1.8}$	$99.2_{\pm 1.8}$
Lung_GSE18842	$50.3_{\pm 1.0}$	$82.1_{\pm 15.9}$	$\mathbf{91.1}_{\pm 10.1}$	$50.0_{\pm 0.0}$	$92.7_{\pm 6.0}$	$90.8_{\pm 9.5}$	$57.2_{\pm 6.9}$	$84.9_{\pm 12.3}$	$\mathbf{88.9}_{\pm 8.3}$	$50.4_{\pm 1.1}$	$94.9_{\pm 5.2}$	$90.8_{\pm 8.7}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Lung_GSE19804	$50.3_{\pm 0.8}$	$78.4_{\pm 7.7}$	$\mathbf{81.4}_{\pm 9.0}$	$50.0_{\pm 0.0}$	$83.8_{\pm 7.2}$	$82.8_{\pm 5.4}$	$58.2_{\pm 7.0}$	$77.8_{\pm 9.1}$	$\mathbf{86.7}_{\pm 3.9}$	$65.5_{\pm 11.8}$	$85.3_{\pm 4.9}$	$79.3_{\pm 11.2}$	$91.8_{\pm 3.2}$	$92.1_{\pm 3.3}$	$91.8_{\pm 3.2}$
Lung_GSE27262	$50.0_{\pm 0.0}$	$75.0_{\pm 16.1}$	$\mathbf{82.6}_{\pm 16.1}$	$50.0_{\pm 0.0}$	$73.6_{\pm 17.6}$	$\mathbf{78.5}_{\pm 18.2}$	$42.4_{\pm 8.2}$	$76.4_{\pm 14.1}$	$\mathbf{77.1}_{\pm 16.4}$	$44.4_{\pm 5.5}$	$80.6_{\pm 12.7}$	$\mathbf{84.7}_{\pm 12.9}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Lung_GSE7670	$50.0_{\pm 0.0}$	$67.7_{\pm 11.3}$	$\mathbf{76.4}_{\pm 17.8}$	$50.0_{\pm 0.0}$	$74.6_{\pm 8.6}$	$67.4_{\pm 14.0}$	$50.0_{\pm 0.0}$	$63.7_{\pm 11.9}$	$\mathbf{69.4}_{\pm 16.6}$	$50.0_{\pm 0.0}$	$73.2_{\pm 16.8}$	$73.0_{\pm 15.6}$	$96.0_{\pm 4.0}$	$96.0_{\pm 4.0}$	$96.0_{\pm 4.0}$
Ovary_GSE6008	$25.0_{\pm 0.0}$	$27.9_{\pm 6.5}$	$27.7_{\pm 3.3}$	$25.6_{\pm 2.4}$	$29.0_{\pm 7.8}$	$28.1_{\pm 6.3}$	$29.4_{\pm 4.3}$	$32.4_{\pm 7.1}$	$25.7_{\pm 2.7}$	$25.0_{\pm 0.0}$	$28.0_{\pm 5.4}$	$26.6_{\pm 2.8}$	$68.6_{\pm 6.2}$	$66.9_{\pm 5.9}$	$\mathbf{69.7}_{\pm 8.1}$
Pancreatic_GSE16515	$50.0_{\pm 0.0}$	$61.6_{\pm 13.1}$	$56.2_{\pm 11.6}$	$50.0_{\pm 0.0}$	$59.2_{\pm 13.3}$	$57.6_{\pm 12.9}$	$50.0_{\pm 0.0}$	$59.4_{\pm 12.3}$	$58.1_{\pm 10.3}$	$50.0_{\pm 0.0}$	$60.7_{\pm 10.4}$	$\mathbf{63.1}_{\pm 15.2}$	$82.9_{\pm 13.0}$	$81.8_{\pm 11.9}$	$\mathbf{82.2}_{\pm 11.6}$
Prostate_GSE26910	$52.8_{\pm 7.9}$	$50.0_{\pm 20.4}$	$\mathbf{58.3}_{\pm 16.7}$	$50.0_{\pm 0.0}$	$44.4_{\pm 15.7}$	$\mathbf{50.0}_{\pm 0.0}$	$58.3_{\pm 31.2}$	$44.4_{\pm 15.7}$	$\mathbf{61.1}_{\pm 17.1}$	$55.6_{\pm 10.4}$	$47.2_{\pm 14.2}$	$\mathbf{58.3}_{\pm 11.8}$	$88.9_{\pm 12.4}$	$86.1_{\pm 17.1}$	$\mathbf{88.9}_{\pm 12.4}$
Prostate_GSE46602	$50.0_{\pm 0.0}$	$56.9_{\pm 14.0}$	$\mathbf{66.3}_{\pm 13.3}$	$50.0_{\pm 0.0}$	$58.5_{\pm 17.3}$	$55.2_{\pm 11.3}$	$50.0_{\pm 0.0}$	$56.3_{\pm 13.4}$	$\mathbf{59.0}_{\pm 13.6}$	$50.0_{\pm 0.0}$	$55.4_{\pm 12.6}$	$\mathbf{60.5}_{\pm 15.8}$	$83.2_{\pm 12.0}$	$84.2_{\pm 10.3}$	$83.2_{\pm 12.0}$
Prostate_GSE55945	$52.8_{\pm 19.3}$	$56.5_{\pm 17.0}$	$41.7_{\pm 11.8}$	$48.6_{\pm 20.8}$	$42.6_{\pm 10.7}$	$\mathbf{44.4}_{\pm 8.6}$	$50.0_{\pm 0.0}$	$47.2_{\pm 8.1}$	$45.8_{\pm 6.2}$	$50.0_{\pm 0.0}$	$51.9_{\pm 16.4}$	$48.6_{\pm 7.1}$	$79.6_{\pm 12.5}$	$77.8_{\pm 13.0}$	$\mathbf{82.4}_{\pm 10.7}$
Prostate_GSE6919_U95Av2	$50.5_{\pm 1.5}$	$59.7_{\pm 4.0}$	$56.7_{\pm 6.8}$	$50.0_{\pm 0.0}$	$57.8_{\pm 6.0}$	$\mathbf{60.0}_{\pm 4.4}$	$52.9_{\pm 3.3}$	$58.0_{\pm 6.7}$	$57.8_{\pm 6.8}$	$55.9_{\pm 7.1}$	$55.3_{\pm 4.3}$	$55.2_{\pm 4.6}$	$63.2_{\pm 4.6}$	$61.7_{\pm 5.9}$	$\mathbf{64.3}_{\pm 4.4}$
Prostate_GSE6919_U95B	$50.3_{\pm 0.7}$	$52.9_{\pm 4.8}$	$\mathbf{54.3}_{\pm 5.1}$	$50.0_{\pm 0.0}$	$51.1_{\pm 5.5}$	$\mathbf{53.6}_{\pm 5.7}$	$50.1_{\pm 0.2}$	$49.6_{\pm 5.9}$	$\mathbf{54.7}_{\pm 6.7}$	$50.0_{\pm 0.0}$	$52.8_{\pm 6.3}$	$\mathbf{54.2}_{\pm 6.9}$	$64.1_{\pm 5.5}$	$64.3_{\pm 5.7}$	$\mathbf{64.3}_{\pm 5.7}$
Renal_GSE53757	$50.0_{\pm 0.0}$	$74.2_{\pm 4.9}$	$\mathbf{78.2}_{\pm 10.9}$	$50.0_{\pm 0.0}$	$74.2_{\pm 10.0}$	$\mathbf{75.8}_{\pm 11.4}$	$45.9_{\pm 2.9}$	$75.2_{\pm 6.2}$	$\mathbf{76.7}_{\pm 6.9}$	$38.1_{\pm 5.0}$	$75.1_{\pm 7.6}$	$69.5_{\pm 11.2}$	$84.6_{\pm 6.0}$	$82.9_{\pm 7.5}$	$\mathbf{84.6}_{\pm 6.0}$
Renal_GSE6344_U133A	$51.9_{\pm 5.2}$	$78.7_{\pm 15.1}$	$64.8_{\pm 16.9}$	$50.0_{\pm 0.0}$	$75.0_{\pm 18.3}$	$62.0_{\pm 14.1}$	$44.4_{\pm 15.7}$	$77.3_{\pm 12.4}$	$63.9_{\pm 30.6}$	$44.4_{\pm 7.9}$	$72.7_{\pm 17.5}$	$62.5_{\pm 15.0}$	$84.3_{\pm 7.0}$	$84.3_{\pm 7.0}$	$84.3_{\pm 7.0}$
Renal_GSE6344_U133B	$54.2_{\pm 6.8}$	$62.0_{\pm 11.9}$	$55.6_{\pm 10.4}$	$50.0_{\pm 0.0}$	$65.7_{\pm 14.9}$	$63.0_{\pm 14.8}$	$51.9_{\pm 5.2}$	$65.3_{\pm 23.1}$	$60.2_{\pm 14.2}$	$53.7_{\pm 6.9}$	$69.0_{\pm 18.5}$	$49.1_{\pm 18.3}$	$79.6_{\pm 6.6}$	$79.6_{\pm 6.6}$	$79.6_{\pm 6.6}$
Renal_GSE66270	$50.0_{\pm 0.0}$	$73.9_{\pm 13.5}$	$\mathbf{89.2}_{\pm 13.9}$	$50.0_{\pm 0.0}$	$73.9_{\pm 13.7}$	$\mathbf{80.6}_{\pm 18.7}$	$45.3_{\pm 9.7}$	$73.6_{\pm 14.0}$	$\mathbf{81.7}_{\pm 16.6}$	$37.5_{\pm 16.2}$	$73.3_{\pm 13.0}$	$\mathbf{75.8}_{\pm 11.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
mean	$\mathbf{44.6}_{\pm 10.9}$	$\mathbf{61.1}_{\pm 12.6}$	$\mathbf{64.2}_{\pm 15.7}$	$\mathbf{44.1}_{\pm 11.0}$	$\mathbf{63.0}_{\pm 14.7}$	$\mathbf{63.2}_{\pm 14.2}$	$\mathbf{44.2}_{\pm 11.7}$	$\mathbf{60.5}_{\pm 13.2}$	$\mathbf{62.9}_{\pm 14.9}$	$\mathbf{43.1}_{\pm 12.2}$	$\mathbf{62.1}_{\pm 14.5}$	$\mathbf{62.6}_{\pm 14.6}$	$\mathbf{86.2}_{\pm 12.6}$	$\mathbf{85.6}_{\pm 13.1}$	$\mathbf{86.3}_{\pm 12.3}$

Table 11: Full results on CUMIDA-42, regime `top25rnd75`: all 42 datasets × five models. Balanced accuracy (×100), mean$_{\pm \text{std}}$ over the nine (seed, fold) cells, learning rate chosen per fold on validation; the *mean* row averages over the 42 datasets. ZS: zero-shot; vFT: vanilla full fine-tuning; KG: KNOWSTFM with the per-model recipe (hard mask with LoRA on the four NanoTabPFN checkpoints, soft gate with LoRA on TabPFN-v3). **Bold**: KG > vFT.

Dataset	small-prior, small			small-prior, base			big-prior, small			big-prior, base			TabPFN-v3		
	ZS	vFT	KG	ZS	vFT	KG	ZS	vFT	KG	ZS	vFT	KG	ZS	vFT	KG
Bladder_GSE31189	51.6$_{\pm 4.8}$	49.9$_{\pm 3.5}$	47.5$_{\pm 7.4}$	49.2$_{\pm 3.0}$	51.1$_{\pm 6.3}$	50.7$_{\pm 7.4}$	50.0$_{\pm 0.0}$	50.3$_{\pm 4.6}$	48.0$_{\pm 5.5}$	50.0$_{\pm 0.0}$	47.8$_{\pm 6.5}$	**50.2**$_{\pm 4.7}$	55.0$_{\pm 12.3}$	57.3$_{\pm 9.9}$	56.1$_{\pm 12.5}$
Brain_GSE15824	41.7$_{\pm 9.1}$	48.4$_{\pm 11.2}$	**50.5**$_{\pm 10.1}$	30.6$_{\pm 8.8}$	43.8$_{\pm 10.4}$	**47.7**$_{\pm 13.3}$	34.3$_{\pm 9.1}$	50.7$_{\pm 11.7}$	44.9$_{\pm 12.6}$	28.0$_{\pm 5.7}$	44.0$_{\pm 9.6}$	41.9$_{\pm 12.3}$	71.5$_{\pm 11.7}$	70.8$_{\pm 11.7}$	**72.5**$_{\pm 12.1}$
Brain_GSE50161	25.0$_{\pm 0.0}$	59.1$_{\pm 9.1}$	**65.2**$_{\pm 8.9}$	25.0$_{\pm 0.0}$	61.2$_{\pm 10.0}$	**66.8**$_{\pm 8.5}$	25.0$_{\pm 0.0}$	59.1$_{\pm 7.0}$	**64.4**$_{\pm 9.8}$	25.0$_{\pm 0.0}$	52.8$_{\pm 16.7}$	**73.5**$_{\pm 8.0}$	92.9$_{\pm 5.6}$	92.9$_{\pm 5.6}$	92.6$_{\pm 6.1}$
Breast_GSE10797	33.3$_{\pm 0.0}$	39.6$_{\pm 5.7}$	**46.0**$_{\pm 12.2}$	33.3$_{\pm 0.0}$	43.2$_{\pm 7.4}$	41.6$_{\pm 5.5}$	33.7$_{\pm 1.2}$	40.0$_{\pm 10.3}$	**45.9**$_{\pm 8.0}$	27.8$_{\pm 6.9}$	43.7$_{\pm 8.2}$	**44.9**$_{\pm 11.4}$	77.8$_{\pm 6.4}$	76.1$_{\pm 8.0}$	**76.5**$_{\pm 6.4}$
Breast_GSE26910	50.0$_{\pm 0.0}$	55.6$_{\pm 15.7}$	**66.7**$_{\pm 16.7}$	50.0$_{\pm 0.0}$	61.1$_{\pm 17.1}$	50.0$_{\pm 20.4}$	55.6$_{\pm 15.7}$	55.6$_{\pm 15.7}$	52.8$_{\pm 7.9}$	50.0$_{\pm 0.0}$	50.0$_{\pm 20.4}$	47.2$_{\pm 7.9}$	88.9$_{\pm 23.9}$	88.9$_{\pm 23.9}$	88.9$_{\pm 23.9}$
Breast_GSE42568	50.0$_{\pm 0.0}$	86.0$_{\pm 10.5}$	77.1$_{\pm 17.6}$	50.0$_{\pm 0.0}$	87.0$_{\pm 10.7}$	78.9$_{\pm 14.8}$	50.0$_{\pm 0.0}$	84.4$_{\pm 10.7}$	80.9$_{\pm 15.0}$	50.0$_{\pm 0.0}$	83.8$_{\pm 12.7}$	75.1$_{\pm 16.7}$	95.8$_{\pm 4.6}$	96.0$_{\pm 4.7}$	95.8$_{\pm 4.6}$
Breast_GSE45827	16.7$_{\pm 0.0}$	75.2$_{\pm 13.2}$	69.4$_{\pm 9.4}$	16.8$_{\pm 0.4}$	83.3$_{\pm 5.0}$	79.9$_{\pm 8.9}$	16.7$_{\pm 0.0}$	79.5$_{\pm 7.1}$	78.5$_{\pm 7.4}$	16.3$_{\pm 0.8}$	84.3$_{\pm 7.5}$	82.0$_{\pm 7.5}$	97.9$_{\pm 1.4}$	97.9$_{\pm 1.4}$	97.9$_{\pm 1.4}$
Breast_GSE7904	33.3$_{\pm 0.0}$	64.7$_{\pm 16.9}$	**68.7**$_{\pm 16.7}$	33.5$_{\pm 2.7}$	60.4$_{\pm 21.8}$	**73.4**$_{\pm 22.7}$	33.3$_{\pm 0.0}$	65.9$_{\pm 15.0}$	**72.9**$_{\pm 9.3}$	22.1$_{\pm 6.1}$	51.4$_{\pm 13.9}$	**73.1**$_{\pm 15.5}$	98.9$_{\pm 2.2}$	98.9$_{\pm 2.2}$	98.9$_{\pm 2.2}$
Colorectal_GSE21510	33.3$_{\pm 0.0}$	85.7$_{\pm 8.1}$	**87.8**$_{\pm 8.7}$	33.3$_{\pm 0.0}$	82.5$_{\pm 11.8}$	81.3$_{\pm 12.6}$	33.3$_{\pm 0.0}$	88.5$_{\pm 2.9}$	83.5$_{\pm 10.1}$	33.3$_{\pm 0.0}$	78.6$_{\pm 17.9}$	**80.0**$_{\pm 9.8}$	99.3$_{\pm 1.7}$	99.3$_{\pm 1.7}$	99.3$_{\pm 1.7}$
Colorectal_GSE32323	50.0$_{\pm 0.0}$	71.1$_{\pm 19.7}$	**79.6**$_{\pm 13.3}$	50.0$_{\pm 0.0}$	83.1$_{\pm 13.2}$	82.2$_{\pm 17.7}$	52.2$_{\pm 10.7}$	66.7$_{\pm 12.8}$	**88.3**$_{\pm 12.1}$	48.3$_{\pm 7.0}$	80.0$_{\pm 14.0}$	78.9$_{\pm 17.4}$	96.9$_{\pm 4.5}$	96.9$_{\pm 4.5}$	96.9$_{\pm 4.5}$
Colorectal_GSE41328	50.0$_{\pm 0.0}$	56.9$_{\pm 7.9}$	**68.5**$_{\pm 18.3}$	50.0$_{\pm 0.0}$	56.0$_{\pm 15.1}$	49.5$_{\pm 12.2}$	55.6$_{\pm 11.9}$	59.3$_{\pm 13.9}$	55.6$_{\pm 11.1}$	54.2$_{\pm 11.8}$	57.4$_{\pm 11.4}$	**63.4**$_{\pm 13.9}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
Colorectal_GSE44861	50.3$_{\pm 0.9}$	72.5$_{\pm 8.0}$	**78.7**$_{\pm 6.6}$	50.0$_{\pm 0.0}$	73.4$_{\pm 8.2}$	**78.4**$_{\pm 7.8}$	51.1$_{\pm 2.3}$	74.3$_{\pm 6.8}$	**76.9**$_{\pm 7.0}$	55.3$_{\pm 4.0}$	71.4$_{\pm 6.8}$	**75.5**$_{\pm 10.3}$	85.4$_{\pm 7.0}$	85.1$_{\pm 7.2}$	**85.4**$_{\pm 7.0}$
Colorectal_GSE77953	26.5$_{\pm 4.8}$	63.3$_{\pm 14.9}$	57.5$_{\pm 14.0}$	27.9$_{\pm 3.4}$	53.2$_{\pm 7.9}$	**60.7**$_{\pm 13.2}$	24.3$_{\pm 6.1}$	59.3$_{\pm 8.0}$	55.8$_{\pm 14.7}$	25.0$_{\pm 0.0}$	54.4$_{\pm 11.0}$	51.5$_{\pm 11.2}$	91.3$_{\pm 4.5}$	92.2$_{\pm 4.0}$	91.7$_{\pm 3.8}$
Colorectal_GSE8671	50.0$_{\pm 0.0}$	91.6$_{\pm 8.3}$	85.2$_{\pm 16.1}$	50.0$_{\pm 0.0}$	89.2$_{\pm 11.6}$	**93.1**$_{\pm 8.9}$	50.5$_{\pm 1.5}$	87.6$_{\pm 13.9}$	**92.9**$_{\pm 7.1}$	46.3$_{\pm 6.3}$	93.6$_{\pm 8.9}$	88.7$_{\pm 9.4}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
Gastric_GSE19826	50.0$_{\pm 0.0}$	47.2$_{\pm 9.8}$	**55.6**$_{\pm 11.9}$	50.0$_{\pm 0.0}$	47.2$_{\pm 11.5}$	45.8$_{\pm 14.4}$	62.5$_{\pm 11.8}$	55.6$_{\pm 10.4}$	**65.3**$_{\pm 14.2}$	50.0$_{\pm 8.3}$	51.4$_{\pm 12.4}$	47.2$_{\pm 5.2}$	73.6$_{\pm 3.9}$	73.6$_{\pm 3.9}$	73.6$_{\pm 3.9}$
Gastric_GSE79973	52.3$_{\pm 6.5}$	65.3$_{\pm 14.6}$	62.0$_{\pm 12.9}$	50.0$_{\pm 0.0}$	64.4$_{\pm 12.9}$	60.6$_{\pm 14.2}$	44.0$_{\pm 8.6}$	66.7$_{\pm 16.1}$	61.6$_{\pm 13.9}$	46.3$_{\pm 10.5}$	64.4$_{\pm 15.7}$	62.5$_{\pm 13.0}$	87.0$_{\pm 8.4}$	85.2$_{\pm 7.1}$	**87.0**$_{\pm 8.4}$
Leukemia_GSE14317	50.0$_{\pm 0.0}$	55.6$_{\pm 8.8}$	**62.0**$_{\pm 18.0}$	50.0$_{\pm 0.0}$	61.1$_{\pm 16.2}$	51.9$_{\pm 5.2}$	50.0$_{\pm 0.0}$	52.8$_{\pm 7.9}$	**55.6**$_{\pm 11.1}$	50.0$_{\pm 0.0}$	55.6$_{\pm 10.4}$	52.8$_{\pm 7.9}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
Leukemia_GSE22529_U133A	50.0$_{\pm 0.0}$	58.6$_{\pm 12.4}$	**66.2**$_{\pm 20.8}$	50.0$_{\pm 0.0}$	64.1$_{\pm 17.0}$	**64.3**$_{\pm 17.1}$	50.0$_{\pm 0.0}$	62.5$_{\pm 17.9}$	**67.1**$_{\pm 18.7}$	50.0$_{\pm 0.0}$	60.6$_{\pm 12.0}$	**68.8**$_{\pm 15.2}$	92.7$_{\pm 7.0}$	92.7$_{\pm 7.0}$	92.7$_{\pm 7.0}$
Leukemia_GSE22529_U133B	50.0$_{\pm 0.0}$	71.6$_{\pm 16.0}$	62.9$_{\pm 14.3}$	50.0$_{\pm 0.0}$	75.3$_{\pm 14.9}$	69.0$_{\pm 15.3}$	50.0$_{\pm 0.0}$	69.6$_{\pm 14.3}$	**72.7**$_{\pm 15.4}$	50.0$_{\pm 0.0}$	70.1$_{\pm 17.3}$	**72.7**$_{\pm 16.2}$	93.7$_{\pm 6.8}$	93.7$_{\pm 6.8}$	93.7$_{\pm 6.8}$
Leukemia_GSE28497	14.3$_{\pm 0.0}$	62.1$_{\pm 6.1}$	**63.6**$_{\pm 6.4}$	14.3$_{\pm 0.0}$	65.9$_{\pm 4.9}$	**69.6**$_{\pm 4.7}$	14.3$_{\pm 0.0}$	62.0$_{\pm 6.0}$	**64.9**$_{\pm 6.2}$	14.3$_{\pm 0.0}$	65.0$_{\pm 7.0}$	**70.2**$_{\pm 5.6}$	80.7$_{\pm 3.4}$	80.5$_{\pm 3.4}$	**80.8**$_{\pm 3.4}$
Leukemia_GSE63270	50.0$_{\pm 0.0}$	86.1$_{\pm 7.1}$	**95.0**$_{\pm 5.0}$	50.0$_{\pm 0.0}$	84.8$_{\pm 6.3}$	**90.8**$_{\pm 7.4}$	50.0$_{\pm 0.0}$	89.1$_{\pm 7.0}$	**93.1**$_{\pm 6.4}$	50.0$_{\pm 0.0}$	87.5$_{\pm 8.0}$	**92.0**$_{\pm 6.6}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
Leukemia_GSE71935	50.0$_{\pm 0.0}$	51.6$_{\pm 9.8}$	**62.5**$_{\pm 18.4}$	50.0$_{\pm 0.0}$	52.8$_{\pm 18.4}$	**65.5**$_{\pm 22.5}$	50.0$_{\pm 0.0}$	51.7$_{\pm 3.7}$	**65.8**$_{\pm 18.1}$	50.0$_{\pm 0.0}$	56.8$_{\pm 14.3}$	56.5$_{\pm 13.3}$	88.9$_{\pm 14.0}$	88.9$_{\pm 14.0}$	**91.2**$_{\pm 14.4}$
Leukemia_GSE9476	20.0$_{\pm 0.0}$	61.5$_{\pm 21.2}$	**71.9**$_{\pm 20.4}$	20.0$_{\pm 0.0}$	77.2$_{\pm 8.5}$	75.6$_{\pm 22.0}$	20.0$_{\pm 0.0}$	74.0$_{\pm 9.6}$	62.0$_{\pm 11.1}$	20.0$_{\pm 0.0}$	74.6$_{\pm 10.6}$	**75.8**$_{\pm 9.7}$	99.0$_{\pm 1.2}$	99.0$_{\pm 1.2}$	98.7$_{\pm 1.6}$
Liver_GSE14520_U133_2	50.0$_{\pm 0.0}$	80.7$_{\pm 11.0}$	**85.9**$_{\pm 11.7}$	50.0$_{\pm 0.0}$	84.0$_{\pm 7.3}$	82.1$_{\pm 10.4}$	45.7$_{\pm 5.9}$	80.9$_{\pm 10.9}$	**84.7**$_{\pm 15.5}$	47.9$_{\pm 5.9}$	77.6$_{\pm 13.4}$	**80.8**$_{\pm 10.8}$	98.5$_{\pm 2.8}$	98.5$_{\pm 2.8}$	98.5$_{\pm 2.8}$
Liver_GSE22405	50.0$_{\pm 0.0}$	61.8$_{\pm 13.0}$	**70.8**$_{\pm 13.8}$	50.0$_{\pm 0.0}$	75.0$_{\pm 10.6}$	**75.7**$_{\pm 9.5}$	41.7$_{\pm 7.8}$	72.9$_{\pm 10.6}$	**79.9**$_{\pm 11.3}$	42.4$_{\pm 4.9}$	75.0$_{\pm 14.1}$	**77.8**$_{\pm 9.8}$	90.3$_{\pm 12.5}$	90.3$_{\pm 12.5}$	**91.0**$_{\pm 10.7}$
Liver_GSE60502	50.0$_{\pm 0.0}$	71.3$_{\pm 12.5}$	**82.4**$_{\pm 10.7}$	50.0$_{\pm 0.0}$	80.6$_{\pm 16.2}$	71.3$_{\pm 18.5}$	33.3$_{\pm 8.8}$	73.1$_{\pm 8.6}$	**82.4**$_{\pm 15.4}$	36.1$_{\pm 10.4}$	85.2$_{\pm 10.2}$	71.3$_{\pm 18.5}$	96.3$_{\pm 4.1}$	96.3$_{\pm 4.1}$	96.3$_{\pm 4.1}$
Liver_GSE62232	50.0$_{\pm 0.0}$	70.9$_{\pm 18.9}$	**73.6**$_{\pm 20.8}$	50.0$_{\pm 0.0}$	66.9$_{\pm 13.8}$	65.0$_{\pm 19.2}$	50.0$_{\pm 0.0}$	61.4$_{\pm 18.0}$	**68.2**$_{\pm 19.9}$	50.0$_{\pm 0.0}$	69.0$_{\pm 15.7}$	**72.8**$_{\pm 20.6}$	99.4$_{\pm 0.9}$	99.4$_{\pm 0.9}$	99.4$_{\pm 0.9}$
Lung_GSE18842	50.0$_{\pm 0.0}$	88.9$_{\pm 14.1}$	**93.3**$_{\pm 6.6}$	50.0$_{\pm 0.0}$	95.1$_{\pm 3.5}$	91.9$_{\pm 6.7}$	56.2$_{\pm 9.3}$	93.4$_{\pm 5.2}$	91.5$_{\pm 7.2}$	50.0$_{\pm 0.0}$	94.9$_{\pm 4.8}$	94.5$_{\pm 4.9}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
Lung_GSE19804	50.0$_{\pm 0.0}$	90.1$_{\pm 6.2}$	88.4$_{\pm 7.2}$	50.0$_{\pm 0.0}$	89.5$_{\pm 5.0}$	88.6$_{\pm 6.9}$	52.2$_{\pm 3.8}$	92.1$_{\pm 3.3}$	81.9$_{\pm 12.8}$	63.1$_{\pm 10.5}$	89.2$_{\pm 6.5}$	87.7$_{\pm 8.3}$	95.3$_{\pm 3.0}$	95.3$_{\pm 3.0}$	95.3$_{\pm 3.0}$
Lung_GSE27262	50.0$_{\pm 0.0}$	77.8$_{\pm 12.5}$	**87.5**$_{\pm 10.2}$	50.0$_{\pm 0.0}$	81.9$_{\pm 13.3}$	**82.6**$_{\pm 18.3}$	38.2$_{\pm 10.4}$	78.5$_{\pm 15.6}$	**79.9**$_{\pm 15.0}$	43.1$_{\pm 12.3}$	78.5$_{\pm 17.0}$	**81.9**$_{\pm 14.9}$	99.3$_{\pm 2.0}$	99.3$_{\pm 2.0}$	99.3$_{\pm 2.0}$
Lung_GSE7670	49.8$_{\pm 2.8}$	77.9$_{\pm 13.2}$	**87.2**$_{\pm 13.0}$	50.0$_{\pm 0.0}$	80.6$_{\pm 8.8}$	79.2$_{\pm 10.6}$	50.0$_{\pm 0.0}$	76.5$_{\pm 7.4}$	**81.2**$_{\pm 10.4}$	50.0$_{\pm 0.0}$	80.6$_{\pm 13.9}$	75.9$_{\pm 10.2}$	94.0$_{\pm 3.0}$	94.0$_{\pm 3.0}$	94.0$_{\pm 3.0}$
Ovary_GSE6008	25.0$_{\pm 0.0}$	38.7$_{\pm 8.3}$	**43.8**$_{\pm 11.9}$	24.7$_{\pm 3.1}$	36.2$_{\pm 10.8}$	**50.0**$_{\pm 12.9}$	24.7$_{\pm 3.1}$	37.1$_{\pm 10.1}$	**49.3**$_{\pm 14.6}$	25.0$_{\pm 0.0}$	31.9$_{\pm 10.2}$	**35.3**$_{\pm 13.0}$	74.0$_{\pm 6.8}$	75.0$_{\pm 6.2}$	73.3$_{\pm 7.2}$
Pancreatic_GSE16515	50.0$_{\pm 0.0}$	67.9$_{\pm 13.3}$	**69.9**$_{\pm 17.4}$	50.0$_{\pm 0.0}$	71.9$_{\pm 13.7}$	67.7$_{\pm 15.3}$	50.0$_{\pm 0.0}$	75.0$_{\pm 13.5}$	65.5$_{\pm 16.8}$	50.0$_{\pm 0.0}$	73.3$_{\pm 10.9}$	68.1$_{\pm 17.8}$	86.3$_{\pm 10.7}$	86.3$_{\pm 10.7}$	86.3$_{\pm 10.7}$
Prostate_GSE26910	52.8$_{\pm 7.9}$	50.0$_{\pm 11.8}$	47.2$_{\pm 7.9}$	50.0$_{\pm 0.0}$	38.9$_{\pm 17.1}$	**50.0**$_{\pm 11.8}$	55.6$_{\pm 15.7}$	47.2$_{\pm 7.9}$	**55.6**$_{\pm 10.4}$	55.6$_{\pm 10.4}$	44.4$_{\pm 19.6}$	**47.2**$_{\pm 7.9}$	86.1$_{\pm 12.4}$	86.1$_{\pm 12.4}$	86.1$_{\pm 12.4}$
Prostate_GSE46602	50.0$_{\pm 0.0}$	58.6$_{\pm 7.3}$	**71.5**$_{\pm 16.0}$	50.0$_{\pm 0.0}$	65.2$_{\pm 20.7}$	**69.3**$_{\pm 15.7}$	50.0$_{\pm 0.0}$	59.5$_{\pm 13.7}$	**76.8**$_{\pm 17.6}$	50.0$_{\pm 0.0}$	67.4$_{\pm 14.1}$	**73.9**$_{\pm 15.5}$	93.2$_{\pm 4.9}$	93.2$_{\pm 4.9}$	**94.4**$_{\pm 5.2}$
Prostate_GSE55945	44.0$_{\pm 9.0}$	49.5$_{\pm 6.9}$	**55.1**$_{\pm 7.3}$	51.4$_{\pm 12.4}$	51.9$_{\pm 17.0}$	**53.2**$_{\pm 12.7}$	50.0$_{\pm 0.0}$	50.9$_{\pm 8.5}$	**56.5**$_{\pm 11.0}$	50.0$_{\pm 0.0}$	53.7$_{\pm 10.5}$	52.3$_{\pm 11.6}$	90.7$_{\pm 12.2}$	90.7$_{\pm 12.2}$	90.7$_{\pm 12.2}$
Prostate_GSE6919_U95Av2	50.0$_{\pm 0.0}$	56.8$_{\pm 6.1}$	**60.6**$_{\pm 6.7}$	50.0$_{\pm 0.0}$	59.0$_{\pm 5.4}$	57.4$_{\pm 6.6}$	53.3$_{\pm 3.9}$	56.3$_{\pm 5.6}$	**61.6**$_{\pm 6.9}$	52.8$_{\pm 7.0}$	58.3$_{\pm 8.1}$	**60.1**$_{\pm 7.8}$	62.6$_{\pm 5.1}$	66.1$_{\pm 4.5}$	63.4$_{\pm 5.1}$
Prostate_GSE6919_U95B	50.0$_{\pm 0.0}$	60.9$_{\pm 3.5}$	**62.7**$_{\pm 7.4}$	50.0$_{\pm 0.0}$	60.2$_{\pm 8.3}$	**63.9**$_{\pm 4.9}$	52.1$_{\pm 5.5}$	60.4$_{\pm 6.1}$	**62.8**$_{\pm 5.7}$	50.0$_{\pm 0.0}$	62.5$_{\pm 7.7}$	61.2$_{\pm 7.9}$	64.5$_{\pm 5.4}$	62.4$_{\pm 7.0}$	**65.8**$_{\pm 5.0}$
Renal_GSE53757	50.0$_{\pm 0.0}$	78.7$_{\pm 8.0}$	**80.1**$_{\pm 6.7}$	50.0$_{\pm 0.0}$	80.6$_{\pm 7.2}$	76.6$_{\pm 8.1}$	47.5$_{\pm 4.1}$	75.0$_{\pm 7.3}$	**81.0**$_{\pm 8.2}$	39.4$_{\pm 5.3}$	75.4$_{\pm 10.3}$	**78.0**$_{\pm 7.5}$	85.0$_{\pm 6.2}$	84.8$_{\pm 6.2}$	84.8$_{\pm 6.1}$
Renal_GSE6344_U133A	58.8$_{\pm 18.0}$	75.0$_{\pm 19.9}$	66.2$_{\pm 19.3}$	50.0$_{\pm 0.0}$	67.1$_{\pm 16.4}$	66.2$_{\pm 24.7}$	42.6$_{\pm 15.9}$	75.5$_{\pm 13.4}$	**78.2**$_{\pm 17.5}$	42.6$_{\pm 11.4}$	66.2$_{\pm 21.7}$	63.9$_{\pm 13.6}$	84.3$_{\pm 9.8}$	85.6$_{\pm 9.2}$	85.6$_{\pm 9.2}$
Renal_GSE6344_U133B	48.6$_{\pm 7.1}$	67.6$_{\pm 17.1}$	56.0$_{\pm 12.3}$	50.0$_{\pm 0.0}$	63.9$_{\pm 14.3}$	54.6$_{\pm 8.9}$	51.4$_{\pm 3.9}$	63.4$_{\pm 17.3}$	57.4$_{\pm 14.5}$	50.0$_{\pm 0.0}$	64.8$_{\pm 19.8}$	62.5$_{\pm 12.4}$	84.3$_{\pm 9.8}$	84.3$_{\pm 9.8}$	84.3$_{\pm 9.8}$
Renal_GSE66270	50.0$_{\pm 0.0}$	71.9$_{\pm 13.9}$	**97.8**$_{\pm 6.3}$	50.0$_{\pm 0.0}$	85.8$_{\pm 17.9}$	82.2$_{\pm 20.6}$	47.8$_{\pm 6.3}$	77.8$_{\pm 19.3}$	**93.3**$_{\pm 8.4}$	43.3$_{\pm 15.6}$	80.0$_{\pm 17.1}$	**81.4**$_{\pm 13.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
mean	**44.7**$_{\pm 10.8}$	**66.1**$_{\pm 13.5}$	**69.8**$_{\pm 13.9}$	**44.3**$_{\pm 10.8}$	**68.0**$_{\pm 15.0}$	**68.0**$_{\pm 13.8}$	**44.0**$_{\pm 11.6}$	**67.0**$_{\pm 13.9}$	**69.9**$_{\pm 13.7}$	**42.9**$_{\pm 11.9}$	**66.8**$_{\pm 15.0}$	**67.9**$_{\pm 14.4}$	**89.3**$_{\pm 11.2}$	**89.4**$_{\pm 11.0}$	**89.5**$_{\pm 11.0}$

Table 12: Full results on CUMIDA-42, regime `top50rnd50`: all 42 datasets × five models. Balanced accuracy (×100), mean$_{\pm \text{std}}$ over the nine (seed, fold) cells, learning rate chosen per fold on validation; the *mean* row averages over the 42 datasets. ZS: zero-shot; vFT: vanilla full fine-tuning; KG: KNOWSTFM with the per-model recipe (hard mask with LoRA on the four NanoTabPFN checkpoints, soft gate with LoRA on TabPFN-v3). **Bold**: KG > vFT.

Dataset	small-prior, small			small-prior, base			big-prior, small			big-prior, base			TabPFN-v3		
	ZS	vFT	KG	ZS	vFT	KG	ZS	vFT	KG	ZS	vFT	KG	ZS	vFT	KG
Bladder_GSE31189	$51.5_{\pm 4.2}$	$52.6_{\pm 5.7}$	$\mathbf{52.9}_{\pm 4.7}$	$49.6_{\pm 4.0}$	$51.0_{\pm 6.5}$	$\mathbf{52.2}_{\pm 6.1}$	$50.0_{\pm 0.0}$	$55.3_{\pm 6.4}$	$55.1_{\pm 7.4}$	$50.0_{\pm 0.0}$	$51.2_{\pm 6.8}$	$\mathbf{52.6}_{\pm 5.5}$	$59.2_{\pm 8.8}$	$57.8_{\pm 8.4}$	$\mathbf{59.6}_{\pm 9.1}$
Brain_GSE15824	$47.9_{\pm 5.9}$	$51.2_{\pm 7.4}$	$51.2_{\pm 7.2}$	$37.0_{\pm 11.2}$	$53.7_{\pm 8.1}$	$51.2_{\pm 12.6}$	$39.6_{\pm 10.6}$	$47.9_{\pm 8.7}$	$44.9_{\pm 12.5}$	$24.1_{\pm 6.2}$	$48.6_{\pm 10.0}$	$\mathbf{52.3}_{\pm 12.4}$	$69.7_{\pm 13.9}$	$70.1_{\pm 14.2}$	$69.0_{\pm 12.7}$
Brain_GSE50161	$25.0_{\pm 0.0}$	$61.6_{\pm 7.5}$	$58.8_{\pm 10.6}$	$25.0_{\pm 0.0}$	$57.9_{\pm 12.1}$	$\mathbf{65.3}_{\pm 12.8}$	$25.0_{\pm 0.0}$	$61.0_{\pm 7.2}$	$\mathbf{73.5}_{\pm 8.0}$	$25.0_{\pm 0.0}$	$64.3_{\pm 6.6}$	$\mathbf{73.4}_{\pm 9.8}$	$90.2_{\pm 6.7}$	$89.8_{\pm 6.3}$	$\mathbf{90.2}_{\pm 6.7}$
Breast_GSE10797	$33.3_{\pm 0.0}$	$43.0_{\pm 12.2}$	$\mathbf{51.1}_{\pm 12.5}$	$33.0_{\pm 1.0}$	$45.3_{\pm 11.7}$	$44.0_{\pm 14.3}$	$33.3_{\pm 1.7}$	$42.7_{\pm 7.2}$	$\mathbf{43.5}_{\pm 9.8}$	$29.7_{\pm 7.6}$	$43.1_{\pm 7.7}$	$\mathbf{45.5}_{\pm 11.5}$	$74.4_{\pm 7.3}$	$74.4_{\pm 7.3}$	$74.0_{\pm 7.0}$
Breast_GSE26910	$50.0_{\pm 0.0}$	$52.8_{\pm 7.9}$	$\mathbf{55.6}_{\pm 19.6}$	$50.0_{\pm 0.0}$	$58.3_{\pm 16.7}$	$55.6_{\pm 19.6}$	$61.1_{\pm 12.4}$	$52.8_{\pm 7.9}$	$\mathbf{61.1}_{\pm 17.1}$	$50.0_{\pm 0.0}$	$58.3_{\pm 16.7}$	$\mathbf{61.1}_{\pm 17.1}$	$91.7_{\pm 16.7}$	$91.7_{\pm 16.7}$	$91.7_{\pm 16.7}$
Breast_GSE42568	$50.0_{\pm 0.0}$	$85.8_{\pm 13.7}$	$82.8_{\pm 13.7}$	$50.0_{\pm 0.0}$	$90.5_{\pm 5.6}$	$87.0_{\pm 6.3}$	$50.0_{\pm 0.0}$	$90.5_{\pm 7.3}$	$89.5_{\pm 6.2}$	$50.0_{\pm 0.0}$	$92.8_{\pm 6.4}$	$87.6_{\pm 9.0}$	$96.3_{\pm 4.9}$	$96.3_{\pm 4.9}$	$96.3_{\pm 4.9}$
Breast_GSE45827	$16.7_{\pm 0.0}$	$73.9_{\pm 18.2}$	$\mathbf{77.4}_{\pm 11.2}$	$16.5_{\pm 0.5}$	$78.3_{\pm 5.5}$	$\mathbf{82.5}_{\pm 5.7}$	$16.7_{\pm 0.0}$	$71.2_{\pm 15.5}$	$\mathbf{80.1}_{\pm 5.3}$	$16.6_{\pm 0.8}$	$79.5_{\pm 6.2}$	$\mathbf{85.1}_{\pm 7.3}$	$97.3_{\pm 1.7}$	$97.6_{\pm 1.3}$	$97.3_{\pm 1.7}$
Breast_GSE7904	$33.3_{\pm 0.0}$	$76.6_{\pm 11.2}$	$75.4_{\pm 15.2}$	$33.3_{\pm 0.0}$	$69.9_{\pm 11.8}$	$\mathbf{73.6}_{\pm 24.8}$	$33.3_{\pm 0.0}$	$75.3_{\pm 13.5}$	$\mathbf{76.5}_{\pm 16.4}$	$24.1_{\pm 9.2}$	$77.1_{\pm 7.8}$	$\mathbf{87.5}_{\pm 10.3}$	$98.3_{\pm 3.3}$	$98.3_{\pm 3.3}$	$98.3_{\pm 3.3}$
Colorectal_GSE21510	$33.3_{\pm 0.0}$	$89.7_{\pm 7.9}$	$86.0_{\pm 11.3}$	$33.3_{\pm 0.0}$	$82.5_{\pm 6.0}$	$\mathbf{83.4}_{\pm 5.1}$	$33.3_{\pm 0.0}$	$87.3_{\pm 9.4}$	$86.1_{\pm 7.0}$	$33.3_{\pm 0.0}$	$76.6_{\pm 16.4}$	$\mathbf{82.6}_{\pm 10.6}$	$99.5_{\pm 1.2}$	$99.5_{\pm 1.2}$	$99.5_{\pm 1.2}$
Colorectal_GSE32323	$50.0_{\pm 0.0}$	$69.6_{\pm 15.6}$	$\mathbf{73.5}_{\pm 20.2}$	$50.0_{\pm 0.0}$	$84.4_{\pm 11.4}$	$78.0_{\pm 11.6}$	$55.9_{\pm 9.3}$	$70.9_{\pm 18.4}$	$\mathbf{74.6}_{\pm 18.8}$	$52.2_{\pm 11.8}$	$73.3_{\pm 18.5}$	$\mathbf{79.1}_{\pm 14.5}$	$96.9_{\pm 4.5}$	$96.9_{\pm 4.5}$	$96.9_{\pm 4.5}$
Colorectal_GSE41328	$50.0_{\pm 0.0}$	$45.4_{\pm 8.9}$	$\mathbf{59.3}_{\pm 19.4}$	$50.0_{\pm 0.0}$	$49.5_{\pm 16.5}$	$48.1_{\pm 12.9}$	$52.8_{\pm 7.9}$	$48.6_{\pm 3.9}$	$\mathbf{56.5}_{\pm 9.4}$	$54.2_{\pm 11.8}$	$54.6_{\pm 17.9}$	$\mathbf{57.9}_{\pm 18.9}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Colorectal_GSE44861	$50.0_{\pm 0.0}$	$73.0_{\pm 12.4}$	$\mathbf{81.2}_{\pm 6.7}$	$50.0_{\pm 0.0}$	$73.5_{\pm 10.3}$	$\mathbf{78.7}_{\pm 8.4}$	$50.1_{\pm 1.4}$	$74.9_{\pm 8.2}$	$\mathbf{81.9}_{\pm 3.8}$	$54.6_{\pm 4.4}$	$73.8_{\pm 8.7}$	$\mathbf{75.6}_{\pm 9.8}$	$86.1_{\pm 5.8}$	$86.1_{\pm 6.1}$	$\mathbf{86.4}_{\pm 5.7}$
Colorectal_GSE77953	$25.5_{\pm 4.1}$	$60.0_{\pm 8.4}$	$\mathbf{61.2}_{\pm 9.6}$	$27.4_{\pm 4.5}$	$63.9_{\pm 16.7}$	$\mathbf{65.4}_{\pm 16.7}$	$25.6_{\pm 3.8}$	$63.0_{\pm 10.3}$	$61.0_{\pm 12.3}$	$25.0_{\pm 0.0}$	$57.7_{\pm 10.9}$	$\mathbf{58.9}_{\pm 14.3}$	$93.1_{\pm 6.9}$	$93.1_{\pm 6.9}$	$93.1_{\pm 6.9}$
Colorectal_GSE8671	$50.0_{\pm 0.0}$	$92.4_{\pm 8.7}$	$\mathbf{96.9}_{\pm 4.8}$	$50.0_{\pm 0.0}$	$93.6_{\pm 6.3}$	$90.9_{\pm 11.0}$	$48.3_{\pm 3.2}$	$90.6_{\pm 6.5}$	$84.1_{\pm 18.2}$	$44.5_{\pm 7.8}$	$89.8_{\pm 14.3}$	$88.4_{\pm 15.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Gastric_GSE19826	$50.0_{\pm 0.0}$	$54.2_{\pm 8.3}$	$\mathbf{61.1}_{\pm 18.1}$	$50.0_{\pm 0.0}$	$50.0_{\pm 13.2}$	$\mathbf{51.4}_{\pm 7.1}$	$54.2_{\pm 14.4}$	$56.9_{\pm 10.4}$	$\mathbf{63.9}_{\pm 10.9}$	$47.2_{\pm 12.9}$	$54.2_{\pm 16.7}$	$\mathbf{58.3}_{\pm 11.8}$	$76.4_{\pm 7.1}$	$76.4_{\pm 7.1}$	$76.4_{\pm 7.1}$
Gastric_GSE79973	$52.3_{\pm 6.5}$	$67.1_{\pm 10.5}$	$60.2_{\pm 14.6}$	$50.0_{\pm 0.0}$	$64.8_{\pm 10.0}$	$\mathbf{65.3}_{\pm 11.8}$	$44.4_{\pm 11.1}$	$65.7_{\pm 11.9}$	$\mathbf{75.9}_{\pm 16.2}$	$46.3_{\pm 6.9}$	$66.2_{\pm 16.9}$	$59.3_{\pm 10.9}$	$90.3_{\pm 7.1}$	$90.3_{\pm 7.1}$	$90.3_{\pm 7.1}$
Leukemia_GSE14317	$50.0_{\pm 0.0}$	$60.2_{\pm 14.0}$	$\mathbf{74.1}_{\pm 19.8}$	$50.0_{\pm 0.0}$	$59.3_{\pm 14.9}$	$\mathbf{71.3}_{\pm 20.1}$	$50.0_{\pm 0.0}$	$58.3_{\pm 14.2}$	$\mathbf{61.1}_{\pm 14.7}$	$50.0_{\pm 0.0}$	$63.0_{\pm 13.7}$	$\mathbf{63.9}_{\pm 17.6}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Leukemia_GSE22529_U133A	$50.0_{\pm 0.0}$	$72.0_{\pm 22.5}$	$64.4_{\pm 20.4}$	$50.0_{\pm 0.0}$	$66.3_{\pm 21.4}$	$65.6_{\pm 21.1}$	$50.0_{\pm 0.0}$	$70.0_{\pm 16.3}$	$\mathbf{75.1}_{\pm 21.8}$	$50.0_{\pm 0.0}$	$64.7_{\pm 19.3}$	$\mathbf{72.7}_{\pm 17.3}$	$92.3_{\pm 7.1}$	$92.3_{\pm 7.1}$	$92.3_{\pm 7.1}$
Leukemia_GSE22529_U133B	$50.0_{\pm 0.0}$	$73.7_{\pm 17.5}$	$\mathbf{74.4}_{\pm 20.6}$	$50.0_{\pm 0.0}$	$73.9_{\pm 20.2}$	$\mathbf{74.6}_{\pm 20.1}$	$50.0_{\pm 0.0}$	$76.6_{\pm 18.7}$	$68.9_{\pm 19.7}$	$50.0_{\pm 0.0}$	$72.9_{\pm 19.5}$	$\mathbf{73.1}_{\pm 18.9}$	$92.9_{\pm 6.7}$	$92.9_{\pm 6.7}$	$92.9_{\pm 6.7}$
Leukemia_GSE28497	$14.2_{\pm 0.2}$	$69.5_{\pm 7.4}$	$68.8_{\pm 5.3}$	$14.3_{\pm 0.0}$	$67.7_{\pm 5.7}$	$\mathbf{74.8}_{\pm 2.7}$	$14.3_{\pm 0.0}$	$69.4_{\pm 3.9}$	$69.2_{\pm 6.2}$	$14.3_{\pm 0.0}$	$66.0_{\pm 7.7}$	$\mathbf{71.7}_{\pm 4.5}$	$81.8_{\pm 2.8}$	$81.5_{\pm 2.6}$	$\mathbf{81.9}_{\pm 2.9}$
Leukemia_GSE63270	$50.0_{\pm 0.0}$	$91.1_{\pm 4.7}$	$\mathbf{95.4}_{\pm 4.8}$	$50.0_{\pm 0.0}$	$92.7_{\pm 4.8}$	$\mathbf{93.8}_{\pm 5.1}$	$50.0_{\pm 0.0}$	$89.8_{\pm 6.7}$	$\mathbf{95.4}_{\pm 4.7}$	$50.0_{\pm 0.0}$	$93.6_{\pm 5.2}$	$\mathbf{94.7}_{\pm 5.3}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Leukemia_GSE71935	$50.0_{\pm 0.0}$	$58.7_{\pm 21.8}$	$\mathbf{66.2}_{\pm 18.1}$	$50.0_{\pm 0.0}$	$65.6_{\pm 16.9}$	$59.8_{\pm 14.6}$	$50.0_{\pm 0.0}$	$57.7_{\pm 17.3}$	$\mathbf{63.4}_{\pm 16.9}$	$50.0_{\pm 0.0}$	$63.2_{\pm 17.9}$	$57.2_{\pm 17.0}$	$97.7_{\pm 6.5}$	$97.7_{\pm 6.5}$	$97.7_{\pm 6.5}$
Leukemia_GSE9476	$20.0_{\pm 0.0}$	$78.1_{\pm 13.7}$	$69.7_{\pm 12.7}$	$20.0_{\pm 0.0}$	$82.5_{\pm 7.4}$	$\mathbf{85.2}_{\pm 6.4}$	$20.0_{\pm 0.0}$	$75.1_{\pm 10.8}$	$67.0_{\pm 17.4}$	$20.0_{\pm 0.0}$	$79.2_{\pm 9.9}$	$75.8_{\pm 15.0}$	$98.7_{\pm 1.6}$	$98.7_{\pm 1.6}$	$\mathbf{99.0}_{\pm 1.5}$
Liver_GSE14520_U133_2	$50.0_{\pm 0.0}$	$87.8_{\pm 10.6}$	$\mathbf{88.6}_{\pm 12.6}$	$50.0_{\pm 0.0}$	$86.2_{\pm 10.1}$	$84.0_{\pm 9.3}$	$44.8_{\pm 5.8}$	$83.1_{\pm 12.2}$	$\mathbf{92.3}_{\pm 6.6}$	$48.6_{\pm 3.9}$	$83.1_{\pm 11.5}$	$82.9_{\pm 14.4}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Liver_GSE22405	$50.0_{\pm 0.0}$	$71.5_{\pm 14.2}$	$\mathbf{75.7}_{\pm 17.3}$	$50.0_{\pm 0.0}$	$69.4_{\pm 15.4}$	$\mathbf{73.6}_{\pm 13.7}$	$36.8_{\pm 11.9}$	$73.6_{\pm 10.1}$	$72.2_{\pm 12.9}$	$38.2_{\pm 11.2}$	$70.1_{\pm 13.1}$	$\mathbf{72.2}_{\pm 15.9}$	$90.3_{\pm 12.5}$	$90.3_{\pm 12.5}$	$90.3_{\pm 12.5}$
Liver_GSE60502	$50.0_{\pm 0.0}$	$75.9_{\pm 14.4}$	$\mathbf{92.6}_{\pm 7.3}$	$50.0_{\pm 0.0}$	$87.0_{\pm 11.9}$	$66.7_{\pm 11.8}$	$33.3_{\pm 9.6}$	$73.1_{\pm 18.3}$	$\mathbf{78.7}_{\pm 13.7}$	$33.3_{\pm 9.6}$	$81.5_{\pm 14.0}$	$62.0_{\pm 15.8}$	$96.3_{\pm 4.1}$	$96.3_{\pm 4.1}$	$96.3_{\pm 4.1}$
Liver_GSE62232	$50.0_{\pm 0.0}$	$73.9_{\pm 16.5}$	$68.9_{\pm 13.2}$	$50.0_{\pm 0.0}$	$71.7_{\pm 16.8}$	$70.3_{\pm 16.0}$	$50.0_{\pm 0.0}$	$72.9_{\pm 19.5}$	$70.7_{\pm 12.9}$	$50.0_{\pm 0.0}$	$75.2_{\pm 14.6}$	$68.4_{\pm 13.5}$	$99.6_{\pm 0.8}$	$99.6_{\pm 0.8}$	$99.6_{\pm 0.8}$
Lung_GSE18842	$50.0_{\pm 0.0}$	$93.2_{\pm 7.1}$	$\mathbf{95.2}_{\pm 4.2}$	$50.0_{\pm 0.0}$	$95.2_{\pm 7.0}$	$94.1_{\pm 8.0}$	$56.9_{\pm 10.3}$	$94.8_{\pm 5.2}$	$92.2_{\pm 7.7}$	$50.4_{\pm 1.1}$	$97.8_{\pm 2.7}$	$\mathbf{98.2}_{\pm 2.8}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
Lung_GSE19804	$50.0_{\pm 0.0}$	$86.7_{\pm 7.9}$	$\mathbf{88.6}_{\pm 7.7}$	$50.0_{\pm 0.0}$	$89.5_{\pm 4.1}$	$85.2_{\pm 6.4}$	$52.7_{\pm 3.5}$	$87.0_{\pm 4.9}$	$\mathbf{89.7}_{\pm 4.0}$	$62.4_{\pm 9.8}$	$83.7_{\pm 15.6}$	$\mathbf{90.1}_{\pm 4.8}$	$95.1_{\pm 2.6}$	$95.1_{\pm 2.6}$	$95.1_{\pm 2.6}$
Lung_GSE27262	$50.0_{\pm 0.0}$	$82.6_{\pm 13.1}$	$\mathbf{86.1}_{\pm 11.7}$	$50.0_{\pm 0.0}$	$84.0_{\pm 10.7}$	$77.1_{\pm 15.3}$	$36.8_{\pm 14.9}$	$75.7_{\pm 15.4}$	$\mathbf{86.1}_{\pm 12.8}$	$35.4_{\pm 9.8}$	$84.0_{\pm 14.2}$	$77.1_{\pm 19.8}$	$99.3_{\pm 2.0}$	$99.3_{\pm 2.0}$	$99.3_{\pm 2.0}$
Lung_GSE7670	$58.0_{\pm 16.8}$	$86.6_{\pm 10.0}$	$\mathbf{87.2}_{\pm 11.8}$	$50.0_{\pm 0.0}$	$84.4_{\pm 13.4}$	$80.3_{\pm 12.5}$	$50.0_{\pm 0.0}$	$81.6_{\pm 8.9}$	$\mathbf{85.6}_{\pm 14.8}$	$50.0_{\pm 0.0}$	$90.6_{\pm 7.9}$	$80.7_{\pm 10.5}$	$95.3_{\pm 2.5}$	$95.3_{\pm 2.5}$	$95.3_{\pm 2.5}$
Ovary_GSE6008	$25.0_{\pm 0.0}$	$38.6_{\pm 8.4}$	$\mathbf{42.2}_{\pm 11.6}$	$25.4_{\pm 4.6}$	$44.6_{\pm 12.3}$	$\mathbf{49.1}_{\pm 7.6}$	$24.8_{\pm 3.3}$	$47.8_{\pm 8.5}$	$38.4_{\pm 8.9}$	$25.0_{\pm 0.0}$	$40.0_{\pm 6.9}$	$\mathbf{42.9}_{\pm 10.4}$	$76.0_{\pm 7.7}$	$76.3_{\pm 9.1}$	$75.7_{\pm 8.1}$
Pancreatic_GSE16515	$50.0_{\pm 0.0}$	$76.3_{\pm 12.5}$	$72.6_{\pm 13.6}$	$50.0_{\pm 0.0}$	$71.0_{\pm 10.4}$	$\mathbf{73.0}_{\pm 7.7}$	$50.0_{\pm 0.0}$	$75.0_{\pm 11.4}$	$74.0_{\pm 10.2}$	$50.0_{\pm 0.0}$	$69.0_{\pm 13.3}$	$\mathbf{76.4}_{\pm 13.2}$	$85.6_{\pm 9.6}$	$85.6_{\pm 9.6}$	$85.6_{\pm 9.6}$
Prostate_GSE26910	$50.0_{\pm 0.0}$	$61.1_{\pm 17.1}$	$61.1_{\pm 12.4}$	$50.0_{\pm 0.0}$	$61.1_{\pm 20.8}$	$52.8_{\pm 7.9}$	$55.6_{\pm 15.7}$	$61.1_{\pm 12.4}$	$55.6_{\pm 15.7}$	$52.8_{\pm 7.9}$	$50.0_{\pm 23.6}$	$\mathbf{55.6}_{\pm 10.4}$	$86.1_{\pm 12.4}$	$86.1_{\pm 12.4}$	$86.1_{\pm 12.4}$
Prostate_GSE46602	$50.0_{\pm 0.0}$	$71.7_{\pm 15.1}$	$\mathbf{78.1}_{\pm 13.0}$	$50.0_{\pm 0.0}$	$77.6_{\pm 11.5}$	$73.4_{\pm 17.2}$	$50.0_{\pm 0.0}$	$72.8_{\pm 14.5}$	$\mathbf{78.3}_{\pm 13.7}$	$50.0_{\pm 0.0}$	$69.2_{\pm 15.0}$	$\mathbf{76.2}_{\pm 15.7}$	$93.8_{\pm 5.3}$	$93.8_{\pm 5.3}$	$93.8_{\pm 5.3}$
Prostate_GSE55945	$49.5_{\pm 6.9}$	$55.1_{\pm 13.0}$	$\mathbf{56.5}_{\pm 18.3}$	$46.8_{\pm 6.1}$	$54.2_{\pm 17.1}$	$52.3_{\pm 8.6}$	$50.0_{\pm 0.0}$	$52.3_{\pm 9.4}$	$\mathbf{52.8}_{\pm 12.9}$	$50.0_{\pm 0.0}$	$53.2_{\pm 14.7}$	$51.4_{\pm 9.2}$	$85.2_{\pm 10.4}$	$85.2_{\pm 10.4}$	$85.2_{\pm 10.4}$
Prostate_GSE6919_U95Av2	$50.0_{\pm 0.0}$	$54.6_{\pm 7.3}$	$\mathbf{60.5}_{\pm 7.7}$	$50.0_{\pm 0.0}$	$59.2_{\pm 8.2}$	$\mathbf{59.5}_{\pm 7.9}$	$51.6_{\pm 3.5}$	$60.6_{\pm 6.0}$	$\mathbf{61.0}_{\pm 6.6}$	$51.6_{\pm 5.8}$	$59.6_{\pm 4.3}$	$\mathbf{62.1}_{\pm 7.3}$	$62.6_{\pm 4.7}$	$64.0_{\pm 3.8}$	$63.1_{\pm 4.4}$
Prostate_GSE6919_U95B	$50.0_{\pm 0.0}$	$65.4_{\pm 4.3}$	$64.8_{\pm 6.8}$	$50.0_{\pm 0.0}$	$64.8_{\pm 6.1}$	$\mathbf{64.9}_{\pm 5.4}$	$56.2_{\pm 5.1}$	$63.5_{\pm 7.8}$	$63.0_{\pm 6.6}$	$50.0_{\pm 0.0}$	$65.0_{\pm 6.9}$	$\mathbf{67.4}_{\pm 6.8}$	$63.4_{\pm 4.8}$	$63.0_{\pm 5.8}$	$\mathbf{63.7}_{\pm 5.1}$
Renal_GSE53757	$50.0_{\pm 0.0}$	$78.0_{\pm 5.7}$	$\mathbf{84.1}_{\pm 5.5}$	$50.0_{\pm 0.0}$	$77.3_{\pm 11.1}$	$75.9_{\pm 7.3}$	$47.4_{\pm 2.9}$	$77.6_{\pm 7.8}$	$\mathbf{81.8}_{\pm 5.3}$	$39.4_{\pm 6.7}$	$77.8_{\pm 9.5}$	$\mathbf{80.1}_{\pm 8.8}$	$85.0_{\pm 6.4}$	$83.4_{\pm 6.2}$	$\mathbf{84.8}_{\pm 6.7}$
Renal_GSE6344_U133A	$50.0_{\pm 13.7}$	$67.6_{\pm 18.2}$	$58.8_{\pm 13.7}$	$50.0_{\pm 0.0}$	$62.5_{\pm 16.3}$	$58.8_{\pm 20.6}$	$42.6_{\pm 15.9}$	$62.0_{\pm 17.9}$	$\mathbf{73.1}_{\pm 18.2}$	$44.4_{\pm 11.1}$	$64.4_{\pm 22.2}$	$58.3_{\pm 17.2}$	$84.3_{\pm 7.8}$	$84.3_{\pm 7.8}$	$84.3_{\pm 7.8}$
Renal_GSE6344_U133B	$49.1_{\pm 4.3}$	$69.0_{\pm 18.6}$	$58.8_{\pm 11.2}$	$50.0_{\pm 0.0}$	$65.7_{\pm 13.7}$	$65.7_{\pm 16.4}$	$50.0_{\pm 0.0}$	$69.4_{\pm 20.5}$	$64.8_{\pm 14.4}$	$48.1_{\pm 5.2}$	$61.6_{\pm 15.2}$	$\mathbf{69.4}_{\pm 14.0}$	$81.5_{\pm 10.8}$	$81.5_{\pm 10.8}$	$81.5_{\pm 10.8}$
Renal_GSE66270	$50.0_{\pm 0.0}$	$82.5_{\pm 16.5}$	$\mathbf{90.0}_{\pm 15.6}$	$50.0_{\pm 0.0}$	$79.2_{\pm 16.5}$	$78.1_{\pm 19.5}$	$46.4_{\pm 5.2}$	$81.7_{\pm 18.9}$	$\mathbf{97.8}_{\pm 6.3}$	$43.1_{\pm 12.6}$	$80.8_{\pm 14.5}$	$79.7_{\pm 19.5}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
mean	$\mathbf{44.9}_{\pm 10.9}$	$\mathbf{69.8}_{\pm 13.9}$	$\mathbf{71.6}_{\pm 14.0}$	$\mathbf{44.3}_{\pm 10.6}$	$\mathbf{70.5}_{\pm 13.8}$	$\mathbf{69.8}_{\pm 13.1}$	$\mathbf{43.9}_{\pm 11.5}$	$\mathbf{69.5}_{\pm 12.9}$	$\mathbf{71.8}_{\pm 14.2}$	$\mathbf{42.5}_{\pm 11.9}$	$\mathbf{69.5}_{\pm 13.9}$	$\mathbf{70.6}_{\pm 13.4}$	$\mathbf{89.6}_{\pm 11.1}$	$\mathbf{89.5}_{\pm 11.1}$	$\mathbf{89.6}_{\pm 11.1}$

Table 13: Full results on CUMIDA-42, regime `top100`: all 42 datasets × five models. Balanced accuracy (×100), mean$_{\pm \text{std}}$ over the nine (seed, fold) cells, learning rate chosen per fold on validation; the *mean* row averages over the 42 datasets. ZS: zero-shot; vFT: vanilla full fine-tuning; KG: KNOWSTFM with the per-model recipe (hard mask with LoRA on the four NanoTabPFN checkpoints, soft gate with LoRA on TabPFN-v3). **Bold**: KG > vFT.

Dataset	small-prior, small			small-prior, base			big-prior, small			big-prior, base			TabPFN-v3		
	ZS	vFT	KG	ZS	vFT	KG	ZS	vFT	KG	ZS	vFT	KG	ZS	vFT	KG
Bladder_GSE31189	51.0$_{\pm 4.5}$	49.5$_{\pm 7.8}$	**53.6**$_{\pm 6.3}$	48.0$_{\pm 3.3}$	51.9$_{\pm 7.2}$	**53.8**$_{\pm 5.3}$	50.0$_{\pm 0.0}$	54.0$_{\pm 6.6}$	52.8$_{\pm 4.3}$	50.0$_{\pm 0.0}$	52.4$_{\pm 6.2}$	**53.5**$_{\pm 6.0}$	62.4$_{\pm 3.4}$	62.2$_{\pm 4.0}$	**63.1**$_{\pm 3.5}$
Brain_GSE15824	50.0$_{\pm 0.0}$	60.9$_{\pm 13.6}$	55.6$_{\pm 13.6}$	40.7$_{\pm 11.4}$	54.6$_{\pm 11.0}$	50.7$_{\pm 14.3}$	41.7$_{\pm 7.8}$	55.6$_{\pm 10.2}$	**56.2**$_{\pm 8.4}$	25.7$_{\pm 8.0}$	50.7$_{\pm 13.3}$	**51.9**$_{\pm 12.6}$	73.8$_{\pm 17.0}$	73.8$_{\pm 17.0}$	73.8$_{\pm 17.0}$
Brain_GSE50161	25.0$_{\pm 0.0}$	67.0$_{\pm 10.7}$	63.3$_{\pm 5.7}$	25.0$_{\pm 0.0}$	68.1$_{\pm 13.0}$	64.0$_{\pm 10.8}$	25.0$_{\pm 0.0}$	72.5$_{\pm 9.0}$	64.1$_{\pm 9.4}$	25.0$_{\pm 0.0}$	72.5$_{\pm 8.3}$	**73.5**$_{\pm 10.7}$	91.7$_{\pm 6.2}$	91.8$_{\pm 6.2}$	**93.1**$_{\pm 5.6}$
Breast_GSE10797	33.3$_{\pm 0.0}$	45.0$_{\pm 15.7}$	**56.3**$_{\pm 17.3}$	33.3$_{\pm 0.0}$	42.7$_{\pm 11.4}$	**45.5**$_{\pm 8.7}$	33.3$_{\pm 0.0}$	38.6$_{\pm 9.7}$	**49.8**$_{\pm 13.0}$	28.9$_{\pm 9.8}$	46.6$_{\pm 10.2}$	**56.5**$_{\pm 10.3}$	80.0$_{\pm 7.6}$	76.8$_{\pm 11.9}$	**80.0**$_{\pm 7.6}$
Breast_GSE26910	50.0$_{\pm 0.0}$	52.8$_{\pm 24.8}$	**61.1**$_{\pm 17.1}$	50.0$_{\pm 0.0}$	47.2$_{\pm 14.2}$	**58.3**$_{\pm 16.7}$	50.0$_{\pm 11.8}$	55.6$_{\pm 22.9}$	55.6$_{\pm 19.6}$	50.0$_{\pm 0.0}$	52.8$_{\pm 7.9}$	**61.1**$_{\pm 23.9}$	88.9$_{\pm 12.4}$	88.9$_{\pm 12.4}$	88.9$_{\pm 12.4}$
Breast_GSE42568	50.0$_{\pm 0.0}$	90.8$_{\pm 7.0}$	87.5$_{\pm 10.3}$	50.0$_{\pm 0.0}$	91.1$_{\pm 3.1}$	87.5$_{\pm 7.8}$	50.0$_{\pm 0.0}$	90.8$_{\pm 5.2}$	87.8$_{\pm 7.9}$	50.0$_{\pm 0.0}$	92.2$_{\pm 6.3}$	90.0$_{\pm 4.7}$	96.5$_{\pm 5.0}$	96.5$_{\pm 5.0}$	96.5$_{\pm 5.0}$
Breast_GSE45827	16.9$_{\pm 0.5}$	82.8$_{\pm 7.4}$	**82.9**$_{\pm 6.5}$	16.7$_{\pm 0.0}$	85.5$_{\pm 4.7}$	**87.4**$_{\pm 5.4}$	16.7$_{\pm 0.0}$	79.4$_{\pm 7.0}$	**81.6**$_{\pm 7.6}$	16.7$_{\pm 0.0}$	81.9$_{\pm 10.1}$	**88.4**$_{\pm 5.4}$	97.9$_{\pm 2.1}$	97.9$_{\pm 2.1}$	97.9$_{\pm 2.1}$
Breast_GSE7904	33.3$_{\pm 0.0}$	85.4$_{\pm 11.1}$	75.1$_{\pm 17.8}$	33.3$_{\pm 0.0}$	85.0$_{\pm 11.9}$	81.7$_{\pm 16.9}$	33.3$_{\pm 0.0}$	79.7$_{\pm 14.7}$	**83.5**$_{\pm 13.2}$	23.0$_{\pm 10.1}$	82.0$_{\pm 14.6}$	79.8$_{\pm 11.3}$	98.2$_{\pm 2.5}$	98.2$_{\pm 2.5}$	**98.8**$_{\pm 2.3}$
Colorectal_GSE21510	33.3$_{\pm 0.0}$	95.5$_{\pm 4.4}$	91.5$_{\pm 8.6}$	33.3$_{\pm 0.0}$	86.3$_{\pm 12.1}$	**90.6**$_{\pm 6.6}$	33.3$_{\pm 0.0}$	92.2$_{\pm 8.0}$	90.5$_{\pm 7.7}$	33.3$_{\pm 0.0}$	95.2$_{\pm 4.8}$	89.6$_{\pm 7.9}$	99.4$_{\pm 1.7}$	99.4$_{\pm 1.7}$	99.4$_{\pm 1.7}$
Colorectal_GSE32323	50.0$_{\pm 0.0}$	76.3$_{\pm 15.4}$	60.6$_{\pm 17.3}$	50.0$_{\pm 0.0}$	82.0$_{\pm 14.6}$	82.0$_{\pm 13.3}$	43.7$_{\pm 8.3}$	73.9$_{\pm 15.8}$	72.8$_{\pm 18.3}$	53.7$_{\pm 11.9}$	73.9$_{\pm 17.5}$	**80.7**$_{\pm 18.9}$	98.0$_{\pm 3.8}$	98.0$_{\pm 3.8}$	98.0$_{\pm 3.8}$
Colorectal_GSE41328	50.0$_{\pm 0.0}$	66.7$_{\pm 20.8}$	60.2$_{\pm 16.6}$	50.0$_{\pm 0.0}$	61.6$_{\pm 21.7}$	52.8$_{\pm 7.9}$	51.4$_{\pm 3.9}$	48.6$_{\pm 3.9}$	**59.7**$_{\pm 18.4}$	45.8$_{\pm 11.8}$	62.5$_{\pm 13.0}$	**63.9**$_{\pm 20.8}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
Colorectal_GSE44861	50.0$_{\pm 0.0}$	77.8$_{\pm 6.3}$	**82.5**$_{\pm 6.9}$	50.0$_{\pm 0.0}$	78.1$_{\pm 7.4}$	**80.4**$_{\pm 4.4}$	51.4$_{\pm 2.1}$	75.1$_{\pm 7.8}$	**82.0**$_{\pm 5.9}$	54.9$_{\pm 5.6}$	79.1$_{\pm 7.1}$	73.7$_{\pm 11.8}$	87.4$_{\pm 5.6}$	85.8$_{\pm 6.0}$	**87.4**$_{\pm 5.6}$
Colorectal_GSE77953	25.3$_{\pm 0.9}$	71.1$_{\pm 9.5}$	68.1$_{\pm 12.2}$	28.9$_{\pm 3.1}$	64.8$_{\pm 9.1}$	63.7$_{\pm 14.3}$	24.2$_{\pm 5.2}$	63.1$_{\pm 9.3}$	**65.7**$_{\pm 11.0}$	25.0$_{\pm 0.0}$	62.9$_{\pm 13.1}$	59.9$_{\pm 13.8}$	94.2$_{\pm 5.0}$	94.2$_{\pm 5.0}$	94.2$_{\pm 5.0}$
Colorectal_GSE8671	50.0$_{\pm 0.0}$	92.8$_{\pm 9.2}$	91.8$_{\pm 12.0}$	50.0$_{\pm 0.0}$	92.4$_{\pm 6.7}$	**92.6**$_{\pm 5.4}$	50.0$_{\pm 0.0}$	93.0$_{\pm 6.5}$	92.9$_{\pm 9.1}$	46.5$_{\pm 6.3}$	95.6$_{\pm 3.7}$	88.9$_{\pm 7.9}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
Gastric_GSE19826	51.4$_{\pm 3.9}$	59.7$_{\pm 19.3}$	**70.8**$_{\pm 15.6}$	50.0$_{\pm 0.0}$	62.5$_{\pm 13.2}$	59.7$_{\pm 15.3}$	54.2$_{\pm 10.2}$	66.7$_{\pm 16.7}$	**68.1**$_{\pm 10.4}$	52.8$_{\pm 12.9}$	52.8$_{\pm 12.9}$	**61.1**$_{\pm 12.4}$	73.6$_{\pm 9.2}$	73.6$_{\pm 9.2}$	73.6$_{\pm 9.2}$
Gastric_GSE79973	52.8$_{\pm 12.9}$	60.6$_{\pm 10.6}$	**76.9**$_{\pm 11.8}$	50.0$_{\pm 0.0}$	69.4$_{\pm 12.4}$	65.7$_{\pm 13.4}$	44.4$_{\pm 11.1}$	64.8$_{\pm 10.4}$	63.0$_{\pm 11.0}$	44.4$_{\pm 11.1}$	64.4$_{\pm 12.3}$	64.4$_{\pm 13.0}$	88.9$_{\pm 6.2}$	88.9$_{\pm 6.2}$	88.9$_{\pm 6.2}$
Leukemia_GSE14317	50.0$_{\pm 0.0}$	64.8$_{\pm 16.6}$	**77.8**$_{\pm 21.2}$	50.0$_{\pm 0.0}$	82.4$_{\pm 15.4}$	72.2$_{\pm 18.8}$	50.0$_{\pm 0.0}$	58.3$_{\pm 12.4}$	**71.3**$_{\pm 20.1}$	50.0$_{\pm 0.0}$	74.1$_{\pm 16.4}$	63.0$_{\pm 16.7}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
Leukemia_GSE22529_U133A	50.0$_{\pm 0.0}$	75.1$_{\pm 21.7}$	71.9$_{\pm 19.8}$	50.0$_{\pm 0.0}$	78.9$_{\pm 18.7}$	78.9$_{\pm 21.0}$	50.0$_{\pm 0.0}$	73.8$_{\pm 17.2}$	**75.2**$_{\pm 20.4}$	50.0$_{\pm 0.0}$	81.9$_{\pm 18.2}$	67.7$_{\pm 19.3}$	92.3$_{\pm 7.1}$	92.3$_{\pm 7.1}$	92.3$_{\pm 7.1}$
Leukemia_GSE22529_U133B	50.0$_{\pm 0.0}$	77.1$_{\pm 17.6}$	70.5$_{\pm 21.0}$	50.0$_{\pm 0.0}$	74.3$_{\pm 17.6}$	74.3$_{\pm 19.4}$	50.0$_{\pm 0.0}$	72.4$_{\pm 19.5}$	69.1$_{\pm 19.4}$	50.0$_{\pm 0.0}$	74.5$_{\pm 19.4}$	71.0$_{\pm 17.9}$	93.3$_{\pm 7.0}$	93.3$_{\pm 7.0}$	93.3$_{\pm 7.0}$
Leukemia_GSE28497	14.3$_{\pm 0.0}$	68.0$_{\pm 4.6}$	**71.2**$_{\pm 3.2}$	14.3$_{\pm 0.0}$	69.6$_{\pm 2.6}$	**70.5**$_{\pm 6.1}$	14.3$_{\pm 0.0}$	68.1$_{\pm 4.9}$	**69.5**$_{\pm 5.7}$	14.3$_{\pm 0.0}$	62.6$_{\pm 17.5}$	**69.1**$_{\pm 5.7}$	82.7$_{\pm 2.6}$	82.5$_{\pm 2.4}$	**82.5**$_{\pm 2.7}$
Leukemia_GSE63270	50.0$_{\pm 0.0}$	94.3$_{\pm 3.5}$	**97.7**$_{\pm 2.5}$	50.0$_{\pm 0.0}$	95.1$_{\pm 3.0}$	92.9$_{\pm 6.4}$	50.0$_{\pm 0.0}$	94.8$_{\pm 3.9}$	**96.0**$_{\pm 3.3}$	50.0$_{\pm 0.0}$	95.1$_{\pm 3.2}$	92.7$_{\pm 6.1}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
Leukemia_GSE71935	50.0$_{\pm 0.0}$	61.4$_{\pm 14.1}$	**65.0**$_{\pm 13.3}$	50.0$_{\pm 0.0}$	66.8$_{\pm 17.2}$	62.3$_{\pm 16.2}$	50.0$_{\pm 0.0}$	59.9$_{\pm 14.4}$	**66.8**$_{\pm 15.4}$	50.0$_{\pm 0.0}$	65.1$_{\pm 11.6}$	63.1$_{\pm 13.7}$	94.9$_{\pm 11.7}$	94.9$_{\pm 11.7}$	94.9$_{\pm 11.7}$
Leukemia_GSE9476	20.0$_{\pm 0.0}$	84.2$_{\pm 8.6}$	81.9$_{\pm 11.6}$	20.0$_{\pm 0.0}$	86.5$_{\pm 10.9}$	72.6$_{\pm 9.3}$	20.0$_{\pm 0.0}$	79.6$_{\pm 12.9}$	73.4$_{\pm 14.9}$	20.0$_{\pm 0.0}$	81.2$_{\pm 12.7}$	66.9$_{\pm 9.6}$	99.5$_{\pm 1.0}$	99.5$_{\pm 1.0}$	99.5$_{\pm 1.0}$
Liver_GSE14520_U133_2	49.3$_{\pm 2.0}$	82.2$_{\pm 12.4}$	**92.8**$_{\pm 9.6}$	50.0$_{\pm 0.0}$	82.5$_{\pm 13.6}$	**84.1**$_{\pm 7.0}$	48.6$_{\pm 3.9}$	81.0$_{\pm 19.0}$	**90.8**$_{\pm 9.7}$	48.6$_{\pm 3.9}$	82.8$_{\pm 13.3}$	**86.8**$_{\pm 9.6}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
Liver_GSE22405	50.0$_{\pm 0.0}$	74.3$_{\pm 15.4}$	**76.4**$_{\pm 15.8}$	50.0$_{\pm 0.0}$	70.1$_{\pm 16.9}$	67.4$_{\pm 14.1}$	37.5$_{\pm 9.3}$	72.2$_{\pm 14.5}$	**76.4**$_{\pm 16.1}$	41.0$_{\pm 8.4}$	68.8$_{\pm 14.1}$	**74.3**$_{\pm 13.3}$	92.4$_{\pm 9.2}$	92.4$_{\pm 9.2}$	92.4$_{\pm 9.2}$
Liver_GSE60502	50.0$_{\pm 0.0}$	88.9$_{\pm 5.6}$	88.9$_{\pm 11.8}$	50.0$_{\pm 0.0}$	86.1$_{\pm 8.8}$	72.2$_{\pm 20.4}$	33.3$_{\pm 11.8}$	79.6$_{\pm 17.6}$	**80.6**$_{\pm 20.8}$	35.2$_{\pm 10.2}$	88.0$_{\pm 10.5}$	83.3$_{\pm 7.9}$	97.2$_{\pm 3.9}$	97.2$_{\pm 3.9}$	97.2$_{\pm 3.9}$
Liver_GSE62232	50.0$_{\pm 0.0}$	78.1$_{\pm 16.8}$	**82.0**$_{\pm 14.1}$	50.0$_{\pm 0.0}$	79.9$_{\pm 16.8}$	73.2$_{\pm 14.1}$	50.0$_{\pm 0.0}$	79.0$_{\pm 15.7}$	68.6$_{\pm 11.2}$	50.0$_{\pm 0.0}$	88.5$_{\pm 13.5}$	75.3$_{\pm 14.1}$	99.4$_{\pm 0.9}$	99.4$_{\pm 0.9}$	99.4$_{\pm 0.9}$
Lung_GSE18842	50.0$_{\pm 0.0}$	96.7$_{\pm 4.6}$	**98.1**$_{\pm 3.6}$	50.0$_{\pm 0.0}$	98.2$_{\pm 2.7}$	97.5$_{\pm 4.0}$	55.4$_{\pm 9.2}$	97.7$_{\pm 2.2}$	96.3$_{\pm 9.4}$	48.9$_{\pm 4.0}$	94.6$_{\pm 7.7}$	**95.3**$_{\pm 6.3}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
Lung_GSE19804	50.0$_{\pm 0.0}$	86.6$_{\pm 8.0}$	**88.1**$_{\pm 7.8}$	50.0$_{\pm 0.0}$	91.5$_{\pm 3.7}$	90.4$_{\pm 3.9}$	54.0$_{\pm 10.4}$	87.7$_{\pm 8.0}$	87.1$_{\pm 7.3}$	57.8$_{\pm 3.1}$	89.6$_{\pm 3.0}$	88.8$_{\pm 7.7}$	95.4$_{\pm 2.4}$	95.4$_{\pm 2.4}$	95.4$_{\pm 2.4}$
Lung_GSE27262	50.0$_{\pm 0.0}$	86.1$_{\pm 16.1}$	**90.3**$_{\pm 12.5}$	50.0$_{\pm 0.0}$	90.3$_{\pm 12.5}$	84.0$_{\pm 17.2}$	47.9$_{\pm 10.6}$	85.4$_{\pm 16.1}$	81.2$_{\pm 21.4}$	33.3$_{\pm 11.8}$	91.7$_{\pm 7.8}$	91.7$_{\pm 15.3}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
Lung_GSE7670	57.7$_{\pm 13.6}$	91.2$_{\pm 7.2}$	87.0$_{\pm 7.5}$	50.0$_{\pm 0.0}$	90.0$_{\pm 2.9}$	85.7$_{\pm 5.9}$	50.0$_{\pm 0.0}$	88.6$_{\pm 5.9}$	82.8$_{\pm 13.4}$	50.0$_{\pm 0.0}$	90.9$_{\pm 3.1}$	89.2$_{\pm 5.8}$	95.3$_{\pm 2.5}$	95.3$_{\pm 2.5}$	95.3$_{\pm 2.5}$
Ovary_GSE6008	25.7$_{\pm 2.0}$	45.7$_{\pm 8.8}$	44.2$_{\pm 10.3}$	26.7$_{\pm 6.2}$	49.0$_{\pm 12.8}$	**54.8**$_{\pm 11.0}$	26.3$_{\pm 2.7}$	48.6$_{\pm 11.3}$	48.5$_{\pm 17.1}$	25.0$_{\pm 0.0}$	53.9$_{\pm 14.5}$	42.1$_{\pm 13.0}$	80.3$_{\pm 4.7}$	78.8$_{\pm 5.7}$	**80.5**$_{\pm 4.9}$
Pancreatic_GSE16515	50.0$_{\pm 0.0}$	75.4$_{\pm 16.6}$	**81.2**$_{\pm 13.7}$	50.0$_{\pm 0.0}$	77.8$_{\pm 12.5}$	75.8$_{\pm 16.6}$	50.0$_{\pm 0.0}$	78.4$_{\pm 14.4}$	76.9$_{\pm 15.5}$	50.0$_{\pm 0.0}$	80.4$_{\pm 10.8}$	75.6$_{\pm 14.0}$	85.6$_{\pm 10.3}$	85.6$_{\pm 10.3}$	85.6$_{\pm 10.3}$
Prostate_GSE26910	50.0$_{\pm 0.0}$	55.6$_{\pm 15.7}$	**63.9**$_{\pm 17.1}$	50.0$_{\pm 0.0}$	44.4$_{\pm 10.4}$	**50.0**$_{\pm 11.8}$	58.3$_{\pm 20.4}$	58.3$_{\pm 11.8}$	58.3$_{\pm 11.8}$	55.6$_{\pm 10.4}$	47.2$_{\pm 14.2}$	**58.3**$_{\pm 11.8}$	86.1$_{\pm 12.4}$	86.1$_{\pm 12.4}$	86.1$_{\pm 12.4}$
Prostate_GSE46602	50.0$_{\pm 0.0}$	79.2$_{\pm 12.9}$	**80.9**$_{\pm 9.7}$	50.0$_{\pm 0.0}$	80.0$_{\pm 12.5}$	79.0$_{\pm 12.2}$	50.0$_{\pm 0.0}$	71.2$_{\pm 13.0}$	**76.1**$_{\pm 11.8}$	50.0$_{\pm 0.0}$	76.8$_{\pm 14.2}$	76.6$_{\pm 15.7}$	93.2$_{\pm 4.9}$	93.2$_{\pm 4.9}$	93.2$_{\pm 4.9}$
Prostate_GSE55945	46.8$_{\pm 10.4}$	53.7$_{\pm 12.3}$	**60.2**$_{\pm 23.5}$	51.4$_{\pm 3.9}$	54.6$_{\pm 13.2}$	52.3$_{\pm 15.4}$	50.0$_{\pm 0.0}$	48.6$_{\pm 17.2}$	**58.8**$_{\pm 22.4}$	50.0$_{\pm 0.0}$	55.6$_{\pm 18.5}$	52.8$_{\pm 11.1}$	91.7$_{\pm 7.6}$	91.7$_{\pm 7.6}$	91.7$_{\pm 7.6}$
Prostate_GSE6919_U95Av2	50.0$_{\pm 0.0}$	66.2$_{\pm 10.5}$	65.3$_{\pm 8.3}$	50.0$_{\pm 0.0}$	57.2$_{\pm 6.8}$	**63.1**$_{\pm 9.7}$	51.4$_{\pm 3.5}$	62.0$_{\pm 5.3}$	**67.0**$_{\pm 6.7}$	51.6$_{\pm 5.9}$	60.1$_{\pm 7.5}$	**60.4**$_{\pm 10.4}$	65.3$_{\pm 4.8}$	63.6$_{\pm 6.0}$	63.4$_{\pm 5.5}$
Prostate_GSE6919_U95B	50.0$_{\pm 0.0}$	62.4$_{\pm 7.4}$	**64.6**$_{\pm 5.0}$	50.0$_{\pm 0.0}$	66.0$_{\pm 6.6}$	64.3$_{\pm 7.5}$	55.1$_{\pm 2.4}$	65.4$_{\pm 6.9}$	64.2$_{\pm 5.4}$	50.0$_{\pm 0.0}$	60.6$_{\pm 7.0}$	**65.3**$_{\pm 6.3}$	65.9$_{\pm 6.0}$	65.1$_{\pm 6.7}$	**65.6**$_{\pm 6.3}$
Renal_GSE53757	50.0$_{\pm 0.0}$	77.6$_{\pm 10.9}$	**80.8**$_{\pm 7.6}$	50.0$_{\pm 0.0}$	81.6$_{\pm 6.3}$	78.2$_{\pm 7.0}$	44.4$_{\pm 9.4}$	79.0$_{\pm 11.8}$	**81.5**$_{\pm 6.7}$	41.3$_{\pm 8.2}$	83.6$_{\pm 5.4}$	81.1$_{\pm 8.0}$	83.6$_{\pm 4.9}$	84.3$_{\pm 6.0}$	83.6$_{\pm 4.9}$
Renal_GSE6344_U133A	56.5$_{\pm 16.0}$	61.6$_{\pm 16.9}$	**65.3**$_{\pm 18.4}$	50.0$_{\pm 0.0}$	71.8$_{\pm 20.5}$	58.3$_{\pm 11.9}$	44.4$_{\pm 15.7}$	63.0$_{\pm 24.8}$	**69.9**$_{\pm 17.1}$	50.0$_{\pm 15.7}$	73.6$_{\pm 18.3}$	69.4$_{\pm 20.5}$	84.3$_{\pm 7.8}$	84.3$_{\pm 7.8}$	84.3$_{\pm 7.8}$
Renal_GSE6344_U133B	52.8$_{\pm 7.9}$	65.7$_{\pm 18.7}$	**66.2**$_{\pm 15.5}$	50.0$_{\pm 0.0}$	58.8$_{\pm 17.8}$	**60.6**$_{\pm 13.3}$	53.7$_{\pm 6.9}$	67.6$_{\pm 14.8}$	**77.3**$_{\pm 10.6}$	50.0$_{\pm 7.9}$	65.7$_{\pm 21.7}$	58.8$_{\pm 12.2}$	80.1$_{\pm 10.7}$	80.1$_{\pm 10.7}$	80.1$_{\pm 10.7}$
Renal_GSE66270	50.0$_{\pm 0.0}$	83.3$_{\pm 11.3}$	**95.6**$_{\pm 8.3}$	50.0$_{\pm 0.0}$	82.2$_{\pm 13.7}$	**85.8**$_{\pm 15.2}$	44.4$_{\pm 10.7}$	87.2$_{\pm 15.8}$	**91.9**$_{\pm 11.7}$	43.3$_{\pm 12.5}$	76.9$_{\pm 15.6}$	**91.1**$_{\pm 10.7}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$	100.0$_{\pm 0.0}$
mean	**45.1**$_{\pm 11.0}$	**73.1**$_{\pm 14.0}$	**74.9**$_{\pm 13.3}$	**44.6**$_{\pm 10.4}$	**73.5**$_{\pm 14.8}$	**71.8**$_{\pm 13.7}$	**43.9**$_{\pm 11.2}$	**71.7**$_{\pm 14.1}$	**73.4**$_{\pm 12.6}$	**42.7**$_{\pm 11.8}$	**73.4**$_{\pm 14.6}$	**72.5**$_{\pm 13.5}$	**90.2**$_{\pm 10.3}$	**90.0**$_{\pm 10.6}$	**90.2**$_{\pm 10.4}$

E LLM PROMPTS AND THE LLM'S ROLE IN EACH METHOD VARIANT

This appendix documents the exact prompts used by the agentic Wikidata mapper and the LLM-direct graph (Section 3.1), and clarifies the LLM's role in the random-graph control of Table 3. All LLM calls use `gemini-3-pro-preview` with `temperature=0.0` and a JSON-only response format.

E.1 AGENTIC WIKIDATA MAPPER, STAGE 1: QUERY PROPOSAL

A single batched LLM call sees the dataset description and all column descriptions at once. The LLM proposes Wikidata search queries; it does not itself emit QIDs at this stage.

```
DATASET: {dataset_desc}

TASK: For each column below, propose 3-7 Wikidata search
queries that retrieve the canonical Wikidata entity for the
column's *concept*.  Use synonyms, expanded abbreviations,
and parent-concept terms.  Use cross-column context (e.g.,
seeing other clinical features helps you expand 'trestbps'
to 'resting blood pressure').

Mark 'value_level:  true' if the *values* of the column are
entities (e.g., a 'cap-shape' column in a mushroom dataset
whose values are 'convex', 'flat', 'knobbed') rather than
the column itself.

Provide a 'type_hint' from this set:  gene | chemical
| disease | measurement | anatomy | clinical_finding |
categorical | demographic | physical_property | other.

COLUMNS: {cols_payload}

Respond with strict JSON, one entry per column:
{"<col_name>":  {"queries":  ["q1", "q2", ...], "type_hint":
"...", "value_level":  <bool>, "rationale":  "..."}}
```

E.2 AGENTIC WIKIDATA MAPPER, STAGE 3: DISAMBIGUATION WITH ABSTENTION

A second batched LLM call sees, for each column, the candidate list returned by the Wikidata REST API (`wbsearchentities`) plus the dataset context. The LLM either selects one QID from the candidate list or abstains. This stage's abstention discipline is the load-bearing difference from a naive "prompt-the-LLM-for-the-QID" baseline

```
DATASET: {dataset_desc}

TASK: For each column, pick the BEST Wikidata QID from
its candidate list, or return empty string if none fit.
Prefer concrete concepts over scientific articles.  Use
cross-column context.  Refuse to invent QIDs not in the
candidate list.

COLUMNS_AND_CANDIDATES: {payload}

Respond with strict JSON:
{"<col_name>":  {"qid":  "<Q123 or empty>", "rationale":
"...", "confidence":  0.0-1.0}}.
```

Stages 2 (Wikidata candidate retrieval) and 4 (SPARQL edge discovery) are deterministic and do not involve the LLM.

E.3 LLM-DIRECT GRAPH (ALTERNATIVE STRUCTURAL PRIOR)

A single LLM call proposes feature–feature edges directly from the dataset and column descriptions, with no Wikidata lookup. The LLM emits edges with a self-reported strength in $[0, 1]$ and a one-sentence rationale; we keep edges with strength ≥ 0.5 and binarise.

```
DATASET: {dataset_desc}

COLUMNS: {cols_payload}

TASK: Propose semantically meaningful feature-feature edges
that capture *useful structural priors* for an attention
mechanism.  Two columns should be connected if their
interaction is plausibly informative for predicting the
target -- for example:  same biological pathway, same body
system, same chemical class, same physical mechanism, known
causal/correlational relationship in the literature.

Use your *implicit domain knowledge* from training.  Do NOT
just connect syntactically similar names -- connect features
that are semantically/causally related.

Each edge should include a one-sentence rationale and a
strength score in [0.5, 1.0] (0.5 = weakly related, 1.0 =
strongly/canonically related).

Be selective:  aim for 1-3 edges per column on average,
prioritising strong edges.  Symmetric undirected edges, no
self-loops.

Respond with strict JSON:
{"edges":  [{"from":  "<col_a>", "to":  "<col_b>",
"relation":  "<short label>", "rationale":  "...",
"strength":  <float>}, ...], "global_notes":  "<1-2
sentences on the overall structure you propose>"}
```

The RANDOM GRAPH control replaces the knowledge-derived adjacency with a graph sampled independently of Wikidata, the LLM, and the ground-truth DAG, using a fixed NumPy random seed. There is no entity linking or semantic information in this arm. All remaining adaptation settings are kept unchanged.

F IMPLEMENTATION

We build KNOWSTFM on two inspectable backbone families. The nanoscale models follow NanoTabPFN (`https://github.com/automl/nanoTabPFN`) and are pretrained from explicit synthetic prior dumps using the interface of the nanoTabPFN/TFM-Playground ecosystem (`https://github.com/automl/TFM-Playground/`); we treat the prior family as part of the model identity and evaluate two prior dumps $\times$ two widths, giving the four checkpoints of Table 5. The frontier model is TabPFN-v3 (53M), adapted through the release fine-tuning trainer. Data handling uses scikit-learn (Pedregosa et al., 2011) and training uses PyTorch (Paszke et al., 2019). For every downstream fold, feature selection, scaling and graph construction are fit on the training split only and frozen in a split manifest, so the test split never influences any of them and every arm in a fold sees an identical graph and an identical feature set. The graph prior S is built from PrimeKG for the CUMIDA gene-expression panels, from the agentic Wikidata mapping of Section 3.1 for general-domain tables, and from the generating DAG for the Bayesian-network benchmarks.

In each adapted feature-attention block the prior enters only as a pre-softmax logit term, never as a change to the value path. A HARD block masks off-graph pairs additively (using -10^4 rather than $-\infty$ so the term is representable in fp16, with the diagonal always unmasked so no row is fully suppressed);

a SOFT block adds a learnable per-head bias $g_h^{(\ell)} S$ with $g_h^{(\ell)} = \sigma(a_h^{(\ell)})\, s_h^{(\ell)}$, initialised to 0.5. The schedule over adapted blocks is fixed per backbone family by Appendix G: (HARD, SOFT, SOFT) for NanoTabPFN, and a soft gate on the column-aggregator blocks for TabPFN-v3. Task adaptation uses LoRA with rank $r = 16$ and $\alpha = 32$ on the query, key, value and output projections and on both feed-forward linears of each adapted block. On TabPFN-v3 the base weights are frozen so that only the LoRA factors and the gates receive gradients; attention is computed under the SDPA MATH backend so that gradients reach the gate parameters, and the gates are placed in their own parameter group at a higher learning rate. On that backbone the graph applies to the column axis only, and Appendix G reports the variant that additionally adapts the in-context blocks. Experiments run on single A100 GPUs under a Python 3.11 virtual environment. Statistical comparisons use paired Wilcoxon tests and AutoRank (`https://github.com/sherbold/autorank`; Herbold (2020)).

F.1 HYPERPARAMETERS

Every arm receives an identical search budget: the learning rate is swept over $\eta \in \{5{\times}10^{-5}, 10^{-4}, 5{\times}10^{-4}, 10^{-3}\}$ and selected *per fold* on the validation split, never on test. All remaining hyperparameters are fixed in advance, so no arm can gain from a wider search: LoRA $r = 16$, $\alpha = 32$, and the per-backbone schedule above. Optimisation uses AdamW (Loshchilov & Hutter, 2019). The nanoscale backbones train for up to 500 steps with early stopping on an inner split (patience 25, evaluated every 10 steps, best state restored), weight decay 10^{-5} and a context fraction of 0.8. TabPFN-v3 trains for up to 20 epochs through the release trainer with early stopping (patience 6), an internal validation split of 0.2, and a learning rate of $5{\times}10^{-2}$ for the gate parameters. CUMIDA-42 uses the `rnd100` regime by default, with `top25rnd75`, `top50rnd50` and `top100` reported in Appendix D; `xdomain13` and the Bayesian-network benchmarks use all columns. Every reported number uses 3 seeds $\times$ 3 stratified folds, with the selected configuration refit on the full training split before scoring the held-out test set.

G ON THE ADAPTATION OF THE MODELS

Table 14 decomposes KNOWSTFM into its two design axes on CUMIDA-42 (`rnd100`): the *mechanism* by which the graph enters attention (none, soft mask, hard mask) and the *regime* of adaptation (full fine-tuning or LoRA, $r{=}16$). The two no-graph cells are the graph-free references. Every cell uses the protocol of Section 4, so all six receive an identical four-point learning-rate search and the columns are directly comparable.

Three regularities follow. **(i) The two families prefer opposite mechanisms.** On the NanoTabPFN checkpoints only the hard mask with LoRA moves above vFT ($+3.1$ on the 1.1M small-prior model, $p{=}0.045$; $+2.4$ on its big-prior counterpart; within noise on the two 3.7M models), while the soft mask sits at or below vFT on all four and significantly below it on the small-prior base model. On TabPFN-v3 this inverts: the soft gate with LoRA is the only cell that improves on vFT ($+0.75$, $p{=}0.004$) and the hard mask is the weakest choice in both regimes. A saturated frontier model tolerates a re-weighting of its attention but not the removal of feature pairs; the small models benefit from the stronger inductive bias. **(ii) Full fine-tuning never wins.** Across all five checkpoints the soft- and hard-full cells lie between -1.2 and $+0.3$ of vFT and none is significant: a graph injected into weights that are all free to move is absorbed during tuning, and the low-rank regime is what preserves it. **(iii) LoRA alone is not the source of the gain.** Graph-free LoRA falls 1.2 to 2.6 points below vFT on every NanoTabPFN checkpoint (significant on both base models) and is a non-significant $+0.4$ on TabPFN-v3, so the nano improvement is not a low-rank regularisation effect.

These results fix the per-model recipe used throughout: hard mask with LoRA for NanoTabPFN, soft mask with LoRA for TabPFN-v3. Because the recipe is chosen on this benchmark, the CUMIDA numbers carry a mild optimism from that choice; Table 3 applies it unchanged to the cross-domain datasets as the out-of-sample check. On TabPFN-v3 we also gated the in-context blocks alongside the column aggregator: this stays above vFT ($+0.57$, $p{=}0.015$) but not above the column-only variant (-0.18), so we report the latter.

Table 14: Mechanism × regime ablation on CUMIDA-42 (`rnd100`). Balanced accuracy ×100, mean$_{\pm\text{std}}$ over the 42 datasets, per-fold leak-free learning-rate selection over an identical four-point grid for every arm. Second line of each block: mean paired difference to vanilla full fine-tuning (vFT). Stars: paired Wilcoxon vs. vFT across datasets ($^{*}p < .05$, $^{**}p < .01$). **Bold**: best adapted arm in the row.

	No graph		Soft mask		Hard mask	
Model	Full (vFT)	LoRA	Full	LoRA	Full	LoRA
NanoTabPFN small-prior, small	$61.1_{\pm 12.6}$ —	$59.9_{\pm 12.2}$ -1.19	$61.2_{\pm 13.7}$ $+0.10$	$60.7_{\pm 12.7}$ -0.39	$61.0_{\pm 12.8}$ -0.14	$\mathbf{64.2}_{\pm 15.7}{}^{*}$ $+3.11$
NanoTabPFN small-prior, base	$63.0_{\pm 14.7}$ —	$61.0_{\pm 13.2}{}^{*}$ -2.04	$62.7_{\pm 13.9}$ -0.36	$60.8_{\pm 12.9}{}^{*}$ -2.27	$62.4_{\pm 14.0}$ -0.65	$\mathbf{63.2}_{\pm 14.2}$ $+0.18$
NanoTabPFN big-prior, small	$60.5_{\pm 13.2}$ —	$59.2_{\pm 13.4}$ -1.26	$60.0_{\pm 13.3}$ -0.50	$59.9_{\pm 13.5}$ -0.65	$60.3_{\pm 13.1}$ -0.22	$\mathbf{62.9}_{\pm 14.9}$ $+2.40$
NanoTabPFN big-prior, base	$62.1_{\pm 14.5}$ —	$59.5_{\pm 13.2}{}^{*}$ -2.63	$61.8_{\pm 13.9}$ -0.35	$60.0_{\pm 13.6}$ -2.10	$60.9_{\pm 14.4}$ -1.17	$\mathbf{62.6}_{\pm 14.6}$ $+0.54$
TabPFN-v3 (53M)	$85.6_{\pm 13.1}$ —	$86.0_{\pm 12.5}$ $+0.40$	$85.9_{\pm 13.1}$ $+0.30$	$\mathbf{86.3}_{\pm 12.3}{}^{**}$ $+0.75$	$84.5_{\pm 13.8}$ -1.07	$85.2_{\pm 13.6}$ -0.40

H PER-DATASET CROSS-DOMAIN RESULTS

Each entry is the mean ± standard deviation of held-out balanced accuracy (×100) over the nine (seed, fold) cells, with the learning rate chosen per fold on the validation split. *LoRA* is the graph-free low-rank update; KNOWSTFM uses the Wikidata-grounded graph and KNOWSTFM (LLM) the LLM-elicited graph, both under the per-model recipe.

Table 15: Per-dataset cross-domain results for NanoTabPFN small-prior, small (1.13 M) on `xdomain13`. **Bold**: mean over datasets.

Dataset	Zero-shot	Vanilla FT	LoRA	KNOWSTFM	KNOWSTFM (LLM)
QSAR-biodeg	$50.0_{\pm 0.0}$	$81.1_{\pm 2.5}$	$82.2_{\pm 1.6}$	$81.5_{\pm 2.6}$	$81.8_{\pm 1.1}$
SPECT	$50.0_{\pm 0.0}$	$72.3_{\pm 5.4}$	$70.7_{\pm 4.5}$	$69.6_{\pm 3.5}$	$71.3_{\pm 5.1}$
automobile	$50.0_{\pm 0.0}$	$62.1_{\pm 7.7}$	$67.3_{\pm 9.7}$	$71.7_{\pm 7.7}$	$69.4_{\pm 9.1}$
cirrhosis	$33.3_{\pm 0.0}$	$50.9_{\pm 2.2}$	$51.6_{\pm 2.0}$	$51.4_{\pm 4.0}$	$51.1_{\pm 3.1}$
dermatology	$16.7_{\pm 0.0}$	$88.8_{\pm 5.9}$	$86.7_{\pm 5.9}$	$88.4_{\pm 4.5}$	$86.8_{\pm 4.3}$
diabetes_pima	$50.0_{\pm 0.0}$	$72.0_{\pm 2.5}$	$70.7_{\pm 2.9}$	$72.0_{\pm 3.0}$	$72.4_{\pm 2.9}$
glass	$15.1_{\pm 1.4}$	$51.5_{\pm 5.5}$	$47.0_{\pm 7.2}$	$54.2_{\pm 6.5}$	$53.8_{\pm 10.3}$
glioma	$50.0_{\pm 0.0}$	$85.2_{\pm 2.3}$	$85.4_{\pm 1.3}$	$85.5_{\pm 2.3}$	$86.2_{\pm 2.0}$
heart-h	$50.0_{\pm 0.0}$	$78.5_{\pm 3.3}$	$77.3_{\pm 3.2}$	$80.1_{\pm 4.7}$	$79.9_{\pm 2.6}$
heart-statlog	$50.0_{\pm 0.0}$	$78.1_{\pm 3.1}$	$78.4_{\pm 4.0}$	$77.7_{\pm 4.3}$	$77.0_{\pm 4.1}$
hepatitis	$50.0_{\pm 0.0}$	$69.9_{\pm 12.6}$	$65.2_{\pm 10.6}$	$63.4_{\pm 11.5}$	$66.8_{\pm 7.7}$
steel-plates-fault	$15.6_{\pm 0.8}$	$69.4_{\pm 9.0}$	$65.1_{\pm 6.7}$	$70.4_{\pm 3.9}$	$67.3_{\pm 6.9}$
stroke	$50.0_{\pm 0.0}$	$50.1_{\pm 0.2}$	$50.0_{\pm 0.0}$	$49.9_{\pm 0.3}$	$50.0_{\pm 0.0}$
mean	$\mathbf{40.8}_{\pm 14.4}$	$\mathbf{70.0}_{\pm 12.5}$	$\mathbf{69.1}_{\pm 12.7}$	$\mathbf{70.4}_{\pm 12.2}$	$\mathbf{70.3}_{\pm 12.0}$

Table 16: Per-dataset cross-domain results for NanoTabPFN small-prior, base (3.72 M) on `xdomain13`. **Bold**: mean over datasets.

Dataset	Zero-shot	Vanilla FT	LoRA	KNOWSTFM	KNOWSTFM (LLM)
QSAR-biodeg	$50.0_{\pm 0.0}$	$82.8_{\pm 1.5}$	$81.8_{\pm 5.1}$	$82.8_{\pm 1.1}$	$82.1_{\pm 2.3}$
SPECT	$50.0_{\pm 0.0}$	$70.9_{\pm 8.6}$	$71.8_{\pm 4.0}$	$70.5_{\pm 5.3}$	$72.3_{\pm 5.9}$
automobile	$50.0_{\pm 0.0}$	$68.3_{\pm 9.5}$	$68.7_{\pm 12.4}$	$69.4_{\pm 11.5}$	$72.6_{\pm 9.1}$
cirrhosis	$33.5_{\pm 0.6}$	$49.1_{\pm 2.5}$	$49.4_{\pm 3.2}$	$51.3_{\pm 3.6}$	$48.9_{\pm 3.9}$
dermatology	$16.7_{\pm 0.0}$	$91.8_{\pm 2.4}$	$89.8_{\pm 4.2}$	$86.5_{\pm 7.1}$	$91.1_{\pm 6.5}$
diabetes_pima	$50.0_{\pm 0.0}$	$69.9_{\pm 4.4}$	$71.2_{\pm 2.0}$	$67.8_{\pm 3.9}$	$70.9_{\pm 2.5}$
glass	$16.3_{\pm 1.3}$	$54.6_{\pm 11.1}$	$50.0_{\pm 8.1}$	$53.0_{\pm 7.9}$	$54.4_{\pm 10.0}$
glioma	$50.0_{\pm 0.0}$	$85.5_{\pm 2.1}$	$85.1_{\pm 2.9}$	$84.7_{\pm 5.8}$	$86.2_{\pm 1.9}$
heart-h	$50.0_{\pm 0.0}$	$78.5_{\pm 2.8}$	$77.6_{\pm 4.9}$	$78.1_{\pm 3.0}$	$79.3_{\pm 4.9}$
heart-statlog	$50.0_{\pm 0.0}$	$79.3_{\pm 3.6}$	$78.3_{\pm 4.2}$	$78.3_{\pm 4.7}$	$77.7_{\pm 4.4}$
hepatitis	$50.0_{\pm 0.0}$	$64.0_{\pm 6.9}$	$68.5_{\pm 8.9}$	$62.7_{\pm 10.7}$	$68.2_{\pm 10.2}$
steel-plates-fault	$14.3_{\pm 0.1}$	$73.1_{\pm 2.5}$	$73.0_{\pm 1.9}$	$72.6_{\pm 2.2}$	$72.2_{\pm 3.2}$
stroke	$50.0_{\pm 0.0}$	$50.1_{\pm 0.2}$	$50.1_{\pm 0.4}$	$50.0_{\pm 0.1}$	$50.0_{\pm 0.0}$
mean	$40.8_{\pm 14.4}$	$70.6_{\pm 12.9}$	$70.4_{\pm 12.8}$	$69.8_{\pm 12.0}$	$\mathbf{71.2}_{\pm 12.7}$

Table 17: Per-dataset cross-domain results for NanoTabPFN big-prior, small (1.13 M) on `xdomain13`. **Bold**: mean over datasets.

Dataset	Zero-shot	Vanilla FT	LoRA	KNOWSTFM	KNOWSTFM (LLM)
QSAR-biodeg	$50.0_{\pm 0.0}$	$81.3_{\pm 1.9}$	$82.3_{\pm 2.3}$	$81.7_{\pm 1.2}$	$81.8_{\pm 1.8}$
SPECT	$50.0_{\pm 0.0}$	$73.9_{\pm 3.6}$	$72.9_{\pm 4.8}$	$71.3_{\pm 4.9}$	$72.8_{\pm 3.7}$
automobile	$50.0_{\pm 0.0}$	$62.7_{\pm 8.9}$	$72.6_{\pm 5.5}$	$69.1_{\pm 6.4}$	$71.1_{\pm 9.4}$
cirrhosis	$33.1_{\pm 0.4}$	$51.0_{\pm 2.8}$	$51.8_{\pm 2.2}$	$52.6_{\pm 3.5}$	$50.3_{\pm 3.6}$
dermatology	$16.7_{\pm 0.0}$	$88.2_{\pm 6.6}$	$88.7_{\pm 6.2}$	$88.4_{\pm 3.7}$	$84.9_{\pm 7.6}$
diabetes_pima	$50.0_{\pm 0.0}$	$69.3_{\pm 2.7}$	$70.7_{\pm 2.7}$	$70.5_{\pm 1.8}$	$70.8_{\pm 2.8}$
glass	$16.0_{\pm 0.6}$	$50.5_{\pm 6.1}$	$48.9_{\pm 9.8}$	$51.1_{\pm 7.9}$	$47.9_{\pm 8.5}$
glioma	$50.0_{\pm 0.0}$	$85.4_{\pm 1.8}$	$86.1_{\pm 1.8}$	$85.8_{\pm 1.3}$	$86.8_{\pm 1.5}$
heart-h	$50.0_{\pm 0.0}$	$79.3_{\pm 2.9}$	$79.7_{\pm 3.2}$	$79.1_{\pm 4.8}$	$77.9_{\pm 3.4}$
heart-statlog	$50.5_{\pm 0.9}$	$77.4_{\pm 3.1}$	$76.7_{\pm 3.0}$	$78.4_{\pm 3.7}$	$76.7_{\pm 4.0}$
hepatitis	$50.0_{\pm 0.0}$	$64.0_{\pm 9.1}$	$65.0_{\pm 9.7}$	$62.4_{\pm 4.3}$	$68.6_{\pm 8.6}$
steel-plates-fault	$12.2_{\pm 1.0}$	$71.5_{\pm 9.0}$	$66.5_{\pm 10.2}$	$71.1_{\pm 3.1}$	$70.7_{\pm 1.7}$
stroke	$50.0_{\pm 0.0}$	$50.0_{\pm 0.1}$	$50.0_{\pm 0.0}$	$50.1_{\pm 0.2}$	$50.0_{\pm 0.0}$
mean	$40.6_{\pm 14.8}$	$69.6_{\pm 12.7}$	$\mathbf{70.1}_{\pm 12.8}$	$\mathbf{70.1}_{\pm 12.4}$	$70.0_{\pm 12.5}$

Table 18: Per-dataset cross-domain results for NanoTabPFN big-prior, base (3.72 M) on `xdomain13`. **Bold**: mean over datasets.

Dataset	Zero-shot	Vanilla FT	LoRA	KNOWSTFM	KNOWSTFM (LLM)
QSAR-biodeg	$50.0_{\pm 0.0}$	$82.3_{\pm 1.4}$	$81.8_{\pm 1.3}$	$82.6_{\pm 2.1}$	$82.2_{\pm 2.3}$
SPECT	$50.0_{\pm 0.0}$	$73.7_{\pm 4.3}$	$70.3_{\pm 4.6}$	$72.5_{\pm 4.2}$	$72.1_{\pm 5.7}$
automobile	$50.0_{\pm 0.0}$	$70.1_{\pm 9.9}$	$65.2_{\pm 7.8}$	$67.6_{\pm 8.5}$	$71.0_{\pm 8.9}$
cirrhosis	$33.3_{\pm 0.0}$	$47.0_{\pm 5.4}$	$48.0_{\pm 2.9}$	$51.2_{\pm 4.9}$	$50.3_{\pm 3.2}$
dermatology	$16.7_{\pm 0.0}$	$91.4_{\pm 3.4}$	$91.2_{\pm 4.5}$	$87.8_{\pm 5.8}$	$88.9_{\pm 6.8}$
diabetes_pima	$50.0_{\pm 0.0}$	$71.7_{\pm 2.2}$	$69.5_{\pm 4.3}$	$71.0_{\pm 3.3}$	$69.9_{\pm 2.7}$
glass	$15.6_{\pm 0.6}$	$53.6_{\pm 8.7}$	$51.1_{\pm 12.5}$	$53.2_{\pm 8.9}$	$57.9_{\pm 8.6}$
glioma	$50.0_{\pm 0.0}$	$85.5_{\pm 1.8}$	$85.1_{\pm 1.8}$	$84.9_{\pm 2.3}$	$85.8_{\pm 2.8}$
heart-h	$50.0_{\pm 0.0}$	$79.4_{\pm 4.1}$	$79.9_{\pm 3.0}$	$78.6_{\pm 4.1}$	$78.7_{\pm 3.8}$
heart-statlog	$50.0_{\pm 0.0}$	$79.1_{\pm 4.9}$	$77.2_{\pm 4.1}$	$78.9_{\pm 3.5}$	$77.5_{\pm 3.3}$
hepatitis	$50.0_{\pm 0.0}$	$65.7_{\pm 9.4}$	$59.6_{\pm 5.6}$	$66.4_{\pm 9.6}$	$69.8_{\pm 7.3}$
steel-plates-fault	$14.3_{\pm 0.0}$	$73.1_{\pm 1.9}$	$72.4_{\pm 3.8}$	$71.7_{\pm 7.0}$	$71.6_{\pm 3.6}$
stroke	$50.0_{\pm 0.0}$	$50.3_{\pm 0.5}$	$50.0_{\pm 0.0}$	$50.0_{\pm 0.1}$	$50.5_{\pm 1.3}$
mean	$40.8_{\pm 14.5}$	$71.0_{\pm 13.1}$	$69.3_{\pm 13.4}$	$70.5_{\pm 12.1}$	$\mathbf{71.2}_{\pm 11.7}$

Table 19: Per-dataset cross-domain results for TabPFN-v3 (53 M) on `xdomain13`. **Bold**: mean over datasets.

Dataset	Zero-shot	Vanilla FT	LoRA	KNOWSTFM	KNOWSTFM (LLM)
QSAR-biodeg	$85.1_{\pm 1.8}$	$85.3_{\pm 1.5}$	$85.1_{\pm 1.6}$	$85.3_{\pm 1.7}$	$85.3_{\pm 1.7}$
SPECT	$67.9_{\pm 6.3}$	$65.3_{\pm 9.5}$	$69.7_{\pm 3.8}$	$69.4_{\pm 6.0}$	$69.1_{\pm 5.7}$
automobile	$84.2_{\pm 7.3}$	$84.7_{\pm 8.1}$	$84.7_{\pm 7.9}$	$84.7_{\pm 7.9}$	$84.3_{\pm 7.4}$
cirrhosis	$51.8_{\pm 2.5}$	$51.8_{\pm 2.8}$	$51.8_{\pm 2.6}$	$51.9_{\pm 2.5}$	$51.8_{\pm 2.6}$
dermatology	$97.3_{\pm 1.5}$	$97.3_{\pm 1.5}$	$97.3_{\pm 1.5}$	$97.3_{\pm 1.5}$	$97.3_{\pm 1.5}$
diabetes_pima	$71.8_{\pm 1.8}$	$69.4_{\pm 7.1}$	$72.4_{\pm 1.9}$	$72.0_{\pm 1.9}$	$72.1_{\pm 2.0}$
glass	$68.6_{\pm 8.4}$	$69.3_{\pm 8.0}$	$70.3_{\pm 7.5}$	$69.2_{\pm 8.8}$	$69.4_{\pm 8.9}$
glioma	$87.5_{\pm 1.4}$	$86.7_{\pm 2.7}$	$87.2_{\pm 2.1}$	$87.4_{\pm 1.2}$	$87.5_{\pm 1.2}$
heart-h	$79.8_{\pm 2.7}$	$79.2_{\pm 4.3}$	$79.6_{\pm 3.9}$	$79.9_{\pm 2.9}$	$80.1_{\pm 2.7}$
heart-statlog	$85.2_{\pm 4.0}$	$79.9_{\pm 11.1}$	$84.7_{\pm 4.0}$	$84.9_{\pm 4.0}$	$85.1_{\pm 4.0}$
hepatitis	$72.4_{\pm 8.0}$	$66.3_{\pm 10.5}$	$73.8_{\pm 8.5}$	$73.4_{\pm 7.7}$	$72.4_{\pm 7.1}$
steel-plates-fault	$84.1_{\pm 1.1}$	$84.0_{\pm 1.2}$	$84.0_{\pm 0.9}$	$84.3_{\pm 1.1}$	$84.4_{\pm 1.0}$
stroke	$50.0_{\pm 0.0}$	$50.0_{\pm 0.0}$	$50.0_{\pm 0.0}$	$50.0_{\pm 0.0}$	$50.0_{\pm 0.0}$
mean	$75.8_{\pm 13.3}$	$74.6_{\pm 13.4}$	$\mathbf{76.2}_{\pm 13.1}$	$76.1_{\pm 13.2}$	$76.1_{\pm 13.3}$

I DATASET-LEVEL CORRELATES OF KNOWSTFM GAINS

The gains from knowledge-informed tuning vary substantially across CUMIDA-42 datasets and checkpoints. To examine whether this heterogeneity can be explained by simple dataset characteristics, we correlate the per-dataset improvement of KNOWSTFM over Vanilla FT with seven properties of each classification task. We focus on the `rnd100` setting, where every dataset is reduced to the same 100 randomly sampled genes, so feature count is constant by construction and cannot explain variation in the gain.

For each dataset and checkpoint, we define the gain as the difference in balanced accuracy between KNOWSTFM and Vanilla FT after averaging over the nine seed–fold evaluations. We compute Spearman correlations across the $N = 42$ datasets with sample size n, number of classes k, samples per class n/k, majority-class fraction, size of the smallest class, class imbalance, and class entropy. The same seven properties are also tested against the corresponding random-graph control, giving 70 tests in total. We therefore apply a Holm correction over the full family of tests.

Table 20: CUMIDA-42 `rnd100`: Spearman correlation between dataset properties and the per-dataset gain of KNOWSTFM over Vanilla FT for each TabPFN-family checkpoint. $N = 42$ datasets, each averaged over 3 seeds $\times$ 3 folds. $^{*}p < 0.05$, $^{**}p < 0.01$ before correction. No association survives Holm correction over the 70 exploratory tests ($p_{\mathrm{adj,min}} = 0.077$); results are therefore descriptive.

Dataset property	small-prior small (1.13M)	big-prior small (1.13M)	small-prior base (3.72M)	big-prior base (3.72M)	TabPFN-v3 (53M)
samples n	$+0.27$	$+0.41^{**}$	$+0.19$	$+0.26$	-0.15
classes k	-0.15	-0.09	$+0.28$	$+0.03$	-0.05
samples per class n/k	$+0.39^{*}$	$+0.49^{**}$	$+0.03$	$+0.28$	-0.10
majority fraction	$+0.25$	$+0.25$	-0.16	$+0.19$	$+0.06$
smallest class	$+0.25$	$+0.27$	-0.14	$+0.03$	-0.17
imbalance $\max / \min$	$+0.14$	$+0.18$	$+0.20$	$+0.30$	$+0.02$
class entropy	-0.14	-0.15	-0.10	-0.27	-0.09

No dataset property reliably predicts the benefit of KNOWSTFM after correction for multiple comparisons. Descriptively, the clearest pattern appears for the two 1.13M checkpoints, where gains tend to increase with the number of samples per class ($\rho = 0.39$ and $\rho = 0.49$). The big-prior 1.13M checkpoint also shows a positive association with total sample size ($\rho = 0.41$). Comparable patterns are weaker for the 3.72M checkpoints and are not apparent for TabPFN-v3.

These results do not support a simple interpretation in which knowledge-informed tuning benefits the most label-scarce datasets. Instead, the descriptive correlations suggest that the smallest checkpoints

may require sufficient supervision to exploit the injected structural prior. This interpretation remains tentative, however, because none of the associations survives the joint multiple-testing correction.

Overall, the analysis reinforces the checkpoint heterogeneity observed in the main experiments but does not identify a single dataset-level property that reliably predicts when knowledge-informed adaptation will be beneficial.

J PER-DATASET WIKIDATA MAPPING STATISTICS ACROSS DATASETS

For each of the 13 cross-domain datasets used in Table 3, we report the number of columns, mapping coverage (the fraction of columns to which the mapper of Section 3.1 assigned a non-empty QID), Wikidata edge counts by hop type (the union pool used in the headline adjacency $S^{(\rho)}$), the LLM-direct edge count, the resulting adjacency densities, and the Jaccard overlap between the two feature–feature edge sets.

Table 21: Mapping coverage and edge statistics for the 13 cross-domain datasets. "cov. %" = fraction of columns to which the agentic mapper assigned a non-empty QID. "direct/1-hop/ancestor" = raw Wikidata edge counts under each relation policy; the headline adjacency takes their union. "dens. %" = fraction of feature pairs connected under the Wikidata-union or LLM-direct adjacency. "Jaccard %" = $100 \cdot |E_{\text{KG}} \cap E_{\text{LLM}}|/|E_{\text{KG}} \cup E_{\text{LLM}}|$ over feature-pair edges. Note: `SPECT` has 0 % KG density despite 1155 raw ancestor edges because all 22 SPECT columns map to the same Wikidata entity (the `F1–F22` columns are abbreviated SPECT-image regions); `QSAR-biodeg` has 0 KG edges because the agentic mapper abstained on 37 of 41 chemical-descriptor columns.

			Wikidata KG (union)				LLM-direct		
Dataset	K	cov. %	direct	1-hop	ancestor	dens. %	$\lvert E\rvert$	dens. %	Jaccard %
`automobile`	25	72	0	4	234	17.3	28	9.3	6.7
`cirrhosis`	17	100	0	0	338	61.0	15	11.0	11.4
`dermatology`	34	38	0	0	178	7.0	25	4.5	8.5
`diabetes_pima`	8	88	0	0	74	64.3	9	32.1	22.7
`glass`	9	100	0	12	173	100.0	9	25.0	25.0
`glioma`	23	100	0	13	980	92.1	24	9.5	10.3
`heart-h`	13	92	0	2	193	56.4	14	17.9	16.0
`heart-statlog`	13	85	0	2	199	55.1	13	16.7	16.7
`hepatitis`	19	95	0	0	522	70.8	19	11.1	9.4
`QSAR-biodeg`	41	7	0	0	0	0.0	31	3.8	0.0
`SPECT`	22	100	0	0	1155	0.0	28	12.1	0.0
`steel-plates-fault`	27	78	0	50	393	8.8	25	7.1	5.7
`stroke`	10	90	0	7	79	46.7	14	31.1	34.6
mean (13)	20	79	0	7	354	44.9	20	14.7	12.8

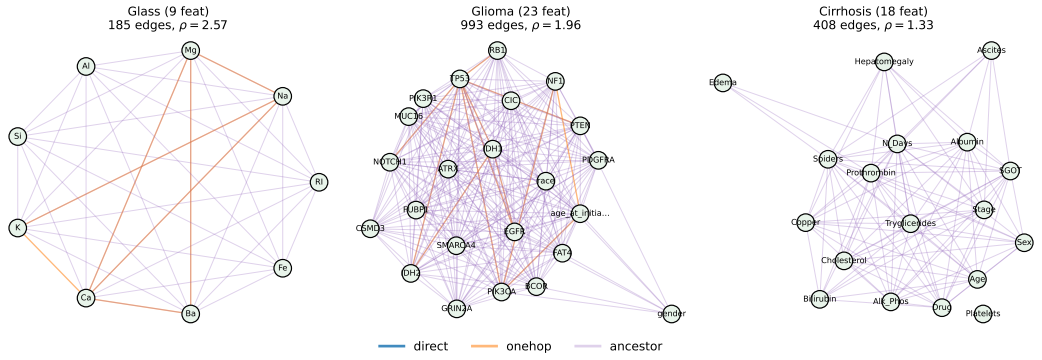


Figure 4: Wikidata adjacency graphs for the three densest XDOMAIN13 datasets (densest first by edges per ordered pair ρ): *Glass* (ρ=2.57), *Glioma* (ρ=1.96), and *Cirrhosis* (ρ=1.33). Edges are coloured by Wikidata relation tier (direct, 1-hop, ancestor); the union over all tiers is the adjacency consumed by the KNOWSTFM injector. Nodes are dataset features placed by Kamad a–Kawai layout. The dominant tier is *ancestor* (ontology parent shared via Wikidata superclass chains), explaining why $\rho > 1$: each pair carries multiple typed edges.

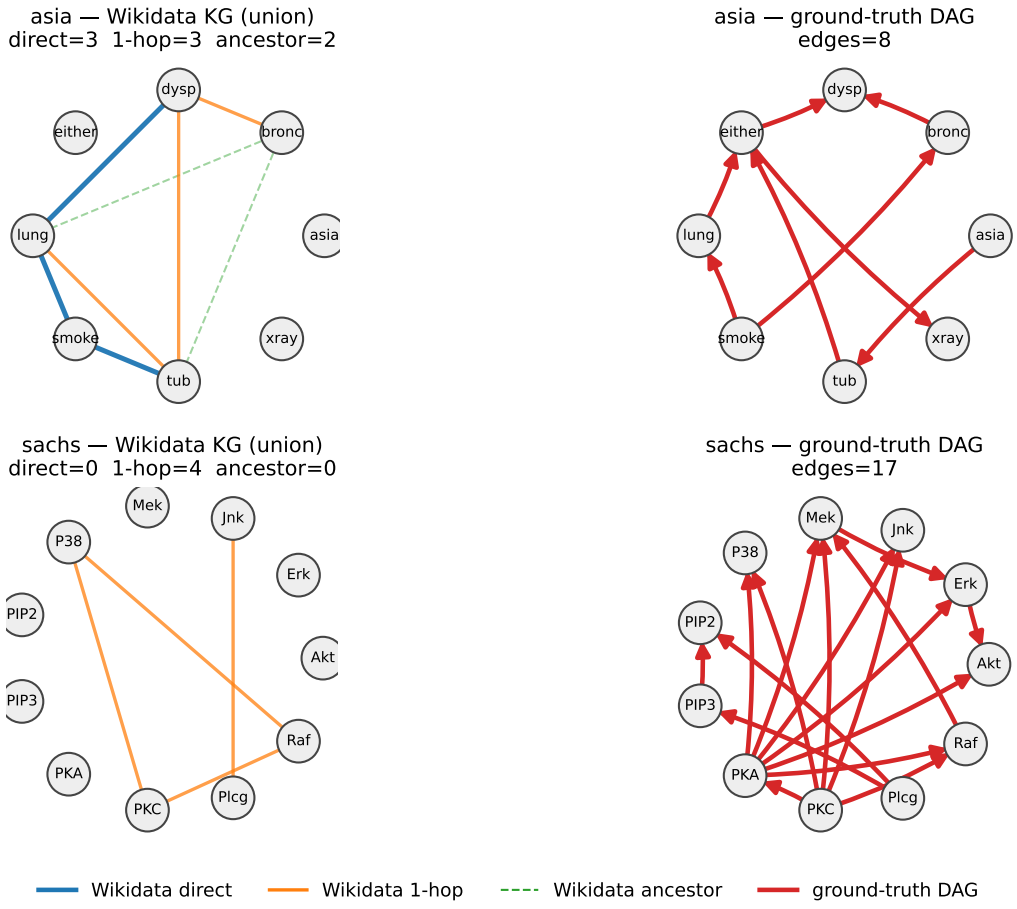
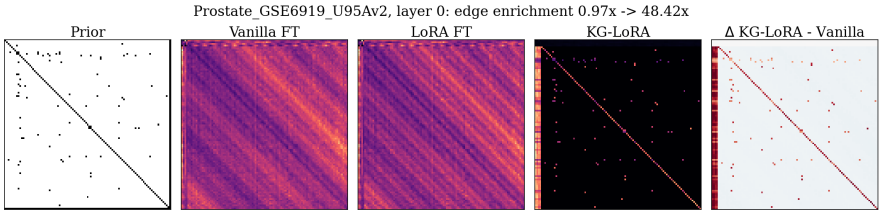


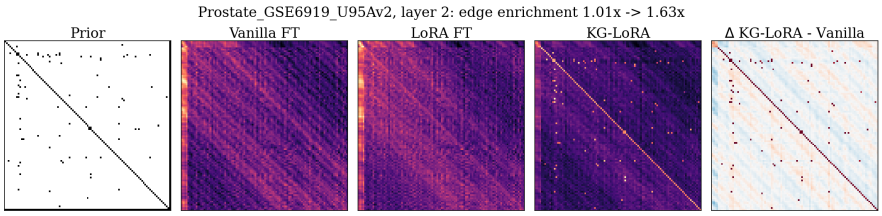
Figure 5: Wikidata-union KG vs. ground-truth DAG on `asia`, `sachs`. Edge colour = Wikidata hop tier (direct / 1-hop / ancestor); red = DAG.

K Additional Attention Visualizations

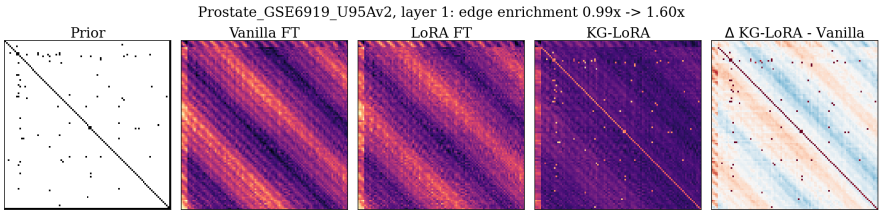
Here we provide layer-wise attention visualizations for the adapted TabPFN model.



(a) Layer 0. The initial hard masking strategy strictly constrains the attention weights, forcing the model to focus purely on the predefined structural priors resulting in a highly sparse pattern.



(b) Layer 1. Transitioning to the first layer of soft masking, the model begins to distribute its attention more broadly, successfully blending the injected structural knowledge with learned semantic context.



(c) Layer 2. Continued soft masking in the deeper layers reveals a highly distributed attention pattern, indicating that the model is synthesizing higher-level, global relationships across the input.

Figure 6: Attention visualization of the TabPFN model on the Prostate-GSE6919 dataset with hard, soft, soft masking.

L TRANSFERRING TO THE TABICL FAMILY

The adaptation recipe in Appendix G was selected on the NanoTabPFN family. To test whether the resulting behaviour is specific to that architecture or extends to nanoscale tabular models more broadly, we repeat the CUMIDA-42 and `xdomain13` protocols on two NanoTabICL checkpoints, pretrained under the small prior. NanoTabICL provides a distinct architecture pretrained under the same controlled synthetic-prior setup. We retain the same data splits and feature panels, the same four-point learning-rate grid selected per fold on validation, and the same three injection variants considered below.

We restrict this transfer experiment to the NanoTabICL implementation for which the feature-attention path can be modified through the same adaptation interface used elsewhere in the paper. We also investigated the larger TabICL variant, but found that its attention mechanism follows a different implementation path and would require architecture-specific modification of internal attention modules to expose the corresponding intervention points. Because this would make the implementation materially different from the controlled protocol used for the other backbones, we do not report results for that variant.

The main qualitative pattern persists. On the headline `rnd100` setting, all three injection variants improve over vanilla fine-tuning: by $+0.77$ to $+1.31$ balanced-accuracy points on the small checkpoint and by $+1.44$ to $+2.65$ points on the base checkpoint, with all three gains significant for the latter. As feature selection becomes stronger, the advantage weakens overall. By `top100`, the effect is essentially neutral on the small checkpoint and negative for all three variants on the base checkpoint. On `xdomain13`, no clear gain remains: every variant lies within 0.4 points of vanilla fine-tuning. Thus, the regime dependence observed in Table 2 is not specific to NanoTabPFN.

Two aspects are architecture-dependent. First, the preferred injection mechanism changes. On NanoTabPFN, the hard mask with LoRA is the only variant that consistently improves over vanilla fine-tuning, whereas the soft gate is neutral or detrimental. On NanoTabICL, by contrast, no clear mechanism preference emerges: the three injection variants produce similar performance and no variant consistently dominates across regimes. This supports selecting the mechanism empirically for each backbone rather than treating the NanoTabPFN recipe as a general rule. Second, the checkpoint that benefits changes. On NanoTabPFN, the gain is concentrated in the 1.1M checkpoints and largely disappears at 3.7M. On NanoTabICL, the larger checkpoint shows the clearest and statistically significant improvement on `rnd100`, while the smaller checkpoint does not. The two families also begin from substantially different zero-shot levels: on `rnd100`, the base NanoTabICL checkpoint reaches 58.9, compared with 44.1 for the corresponding NanoTabPFN checkpoint. The architectures therefore appear to occupy different points on their adaptation curves.

Table 22: Transfer to the NanoTabICL family. Balanced accuracy $\times 100$, averaged over datasets (3 seeds $\times$ 3 folds; learning rate selected per fold on validation). Parentheses report the mean difference Δ from vanilla fine-tuning (vFT). Stars denote paired Wilcoxon tests against vFT ($^{*}p < .05$, $^{**}p < .01$). **Bold** indicates the best result in each row. Gains are strongest in the `rnd100` setting and disappear under fully relevance-selected features or cross-domain transfer, while no single injection mechanism consistently dominates.

Checkpoint	Benchmark	ZS	vFT	soft gate + LoRA	soft gate + full FT	hard mask + LoRA
NanoTabICL small	`rnd100`	44.7	64.5	65.2 (+0.77)	65.8 (+1.30)	**65.8** (+1.31)
	`top25rnd75`	45.7	71.9	71.2 (−0.67)	**72.2** (+0.32)	72.0 (+0.10)
	`top50rnd50`	47.8	74.4	73.6 (−0.84)	73.8 (−0.63)	**75.1** (+0.67)
	`top100`	49.9	77.0	76.9 (−0.09)	**77.3** (+0.31)	76.9 (−0.15)
	`xdomain13`	41.0	72.5	72.4 (−0.13)	**72.8** (+0.25)	72.6 (+0.11)
NanoTabICL base	`rnd100`	58.9	68.6	70.8 (+2.19)*	70.1 (+1.44)*	**71.3** (+2.65)**
	`top25rnd75`	62.0	75.7	76.1 (+0.33)	**76.9** (+1.22)	76.5 (+0.73)
	`top50rnd50`	64.2	77.7	79.0 (+1.30)	78.9 (+1.18)	**79.1** (+1.37)
	`top100`	66.2	**81.9**	80.6 (−1.25)	81.2 (−0.64)	80.9 (−0.98)
	`xdomain13`	43.6	72.8	73.0 (+0.29)	72.7 (−0.02)	**73.1** (+0.38)