TabPFN-Wide: Continued Pre-Training for Extreme Feature Counts

Christopher Kolberg, Katharina Eggensperger & Nico Pfeifer
Department of Computer Science, University of Tübingen
{katharina.eggensperger, nico.pfeifer}@uni-tuebingen.de

Abstract

Revealing novel insights from the relationship between molecular measurements and pathology remains a very impactful application of machine learning in biomedicine. Data in this domain typically contain only a few observations but thousands of potentially noisy features, posing challenges for conventional machine learning approaches. While prior-data fitted networks emerge as foundation models for tabular data, they are currently not suited to handle large feature counts (> 500). Although feature reduction enables their application, it hinders feature importance analysis. We propose a strategy that extends existing models through continued pre-training on synthetic data sampled from a customized prior. The resulting model, TabPFN-Wide, matches or exceeds TabPFNv2's performance while exhibiting improved robustness to noise. It seamlessly scales beyond $50{,}000$ features, regardless of noise levels, while maintaining inherent interpretability, which is critical for biomedical applications. Our results show that prior-informed adaptation is suitable to enhance the capability of foundation models for high-dimensional data.

1 Introduction

Data stored in a table are an important data modality used for quantitative research in healthcare, finance, natural sciences, and many more. Tabular data are relevant for many real-world applications and "offer[s] uniquely exciting, large, unsolved challenges for researchers" [van Breugel and van der Schaar, 2024]. One such challenge is high-dimensional, low-sample-size (HDLSS) data, for example, found in biomedical research. Cohort sizes of studies are small due to cost, time, or disease rarity, while modern biomedical technologies, on the other hand, enable the measurement of thousands of features per patient. Collected data can then be examined, for example, to study interactions between thousands of biomarkers and cancer types [McLendon et al., 2008, Bell et al., 2011].

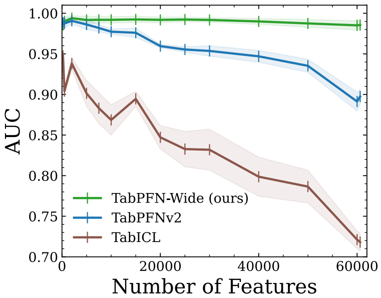


Figure 1: The performance of existing tabular foundation models decreases for a selected high-dimensional biomedical dataset. Further datasets are presented in Section 4 to confirm generality.

Foundation models for structured data have emerged, and models like TabPFN and TabICL [Hollmann et al., 2023b, 2025, Qu et al., 2025] are currently at the forefront of predictive tabular ML benchmark tasks [Erickson et al., 2025]. These state-of-the-art models use in-context learning (ICL) [Brown et al., 2020] and are based on transformers, pre-trained on synthetic or real-world data to solve regression and classification tasks. As a result, they are highly effective on unseen tasks with characteristics similar to those seen during pre-training. While the exact training

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Tabular Data workshop at EurIPS 2025.

data are often unknown, empirical performance on HDLSS data (see brown and blue lines in the example in Figure 1) suggests that current models have not learned to handle extreme feature counts.

Such limits stem from insufficient exposure during pre-training and not necessarily from a lack of model capacity, data or resources; thus, re-training from scratch could be a solution. However, re-training from scratch whenever we encounter a new task or data characteristic to “fix" a model would be extremely resource-intensive, and therefore often infeasible. Instead, we study the more general question: “How can continued pre-training extend tabular foundation models to generalize across diverse task types in high-dimensional, small-sample data?”

Specifically, our contributions are:

1. We develop a novel prior to efficiently generate synthetic HDLSS data.
2. We propose continued pre-training to extend TabPFNv2, resulting in TabPFN-Wide, to handle extreme feature counts beyond $50{,}000$ features.
3. In empirical evaluations on biomedical data and tabular benchmark tasks, we show that TabPFN-Wide maintains performance for small datasets, while being significantly more robust on wide data.

2 Background on Extending Tabular Foundation Models

To address limitations of foundation models like TabPFNv2, current research focuses on scaling to large samples and feature counts. One prominent example is TabICL [Qu et al., 2025], which uses only a fixed number of embedded [CLS] tokens per sample for ICL rather than all the features. Other approaches designed to handle more samples include TuneTables [Feuer et al., 2024] or TabFlex [Zeng et al., 2025]. While all these approaches aim to extend the application range, they propose new architectures and inference mechanisms, often applying feature reduction and compression. In contrast, we aim to expand an *existing* model’s capability without the need for feature reduction.

Continued pre-training as an alternative to fine-tuning has been shown to improve performance on specific downstream tasks. For example, Real-TabPFN [Garg et al., 2025], further pre-trained on real-world datasets, shows significant improvements on real-world tabular benchmarks. We follow this direction, but instead of using real-world data, we study how to continue pre-training with synthetic data to scale TabPFN to extreme feature counts, far beyond what it has seen during pre-training. Because this involves sequential training, it is crucial to prevent the model from experiencing catastrophic forgetting [French, 1993, Kemker et al., 2018]. This could cause the model to perform significantly worse on tabular data within the original ranges of TabPFNv2.

3 Methodology

A Prior for Synthetic HDLSS Data Generation. To adapt our model, we need a mechanism to generate training data, which (1) works fast and cost-effectively, since we need multiple datasets per batch step, and (2) yields realistic data, to provide a meaningful and reliable signal during adaptation.

For the first desideratum, we follow prior work and rely on synthetic data obtained from a data-generating mechanism based on structural causal models [Hollmann et al., 2023a,b]. Datasets are therefore drawn from randomly sampled directed acyclic graphs. Specifically, as the TabPFNv2 prior is not publicly available, we use the open-source prior used to train TabICL [Qu et al., 2025], considering TabICL’s strong empirical performance as evidence of the prior’s similar effectiveness. To fulfill the second desideratum, we leverage the assumption that features in HDLSS data often exhibit substantial noise and strong inter-feature correlations [Clarke et al., 2008].

Specifically, we describe our procedure in Algorithm 1, which takes as input the continuous numerical features X of a dataset, during training sampled from the TabICL prior with a moderate feature count m (generation step), and then artificially *widens* it to $d \gg m$ dimensions (widening step).

Continued Pre-Training. For our continued pre-training setup we start with the original TabPFNv2 classifier checkpoint[1] and updated all parameters during training. We used AdamW (using a weight

[1] See Hugging Face model; Runtime complexity remains unaffected, thus, to satisfy higher resource demands for continued pre-training we used 4 NVIDIA H100 GPUs with a combined memory of 320GB.

decay of 1×10^{-4} and a learning rate of 1×10^{-5}) [Loshchilov and Hutter, 2019] with linear warm-up, cosine decay, and gradient norms clipping to 1.0. We used a batch size of 16, reducing it to 8 for training runs with over 5,000 features due to memory constraints. Training and validation were performed using cross-entropy loss. The generated datasets of the TabICL prior had up to 10 classes (to match TabPFNv2's limitations), 40 to 400 samples, and 50 to 350 features which we then widened using Algorithm 1. The number of features as parameter of Algorithm 1 was uniformly sampled between 200 and d features with $d \in \{1{,}500; 5{,}000; 8{,}000\}$. With a probability of 0.5, the original features were appended to the final dataset. Sparsity and noise level were uniformly sampled with $p \in [0, 0.05]$ and $\sigma \in [0, 1]$ following our analysis visualized in Figure 4. We denote the resulting models as TabPFN-Wide-d, where d indicates the maximum number of features used during training. For model selection, we used two real-world datasets (*COAD* and *GBM*; see description below). We use the model with the lowest average validation loss.

4 Experiments and Results

Datasets and Evaluation Protocol. We use six machine learning–ready TCGA datasets differing from raw TCGA data by already being normalized, quality-checked, and otherwise pre-processed. We use two of them for model selection and the remaining for evaluation (see Section A.1 for further details). In addition, we also evaluate on 21 benchmark tasks (with $\leq 10,000$ samples and $\leq$ 500 features) introduced by *TabArena* [Erickson et al., 2020]. Alongside the foundation models TabPFNv2 and TabICL, we evaluate RealMLP-TD [Holzmüller et al., 2025] as well as classical tree-based machine learning techniques like random forest and XGBoost [Chen and Guestrin, 2016]. Ensembling was not used for TabPFN-Wide, TabPFNv2, TabICL, and RealMLP-TD.

Algorithm 1 Feature Widening

```
Input: Input features $X \in \mathbb{R}^{n \times m}$, target dimension $d$,
       sparsity $p \in [0, 1]$, noise std. $\sigma$
Output: Wide features $X_{wide} \in \mathbb{R}^{n \times d}$
Sample weights $W \in \mathbb{R}^{m \times d}$ with $W_{ij} \sim \mathcal{N}(0, 1)$
Sample mask $M \in \{0, 1\}^{m \times d}$
    with $M_{ij} \sim$ Bernoulli($p$)
Compute wide features $X_{wide} \leftarrow X(M \odot W)$
Sample noise $N \in \mathbb{R}^{m \times d}$
    with $N_{ij} \sim \mathcal{N}(0, (\sigma \sigma_j)^2)$ and $\sigma_j = \text{std}(X_{wide_{:,j}})$
Add noise $X_{wide} \leftarrow X_{wide} + N$
return $X_{wide}$
```

We perform 5-fold cross-validation for our biomedical datasets to compute AUROC, AUPRC, and accuracy. For the TabArena datasets we follow the original evaluation protocol and compute AUROC using a 3-fold cross-validation repeated 3 or 10 times, depending on dataset size.

Results on Real-World Wide Datasets. We first evaluate TabPFN-Wide on the four multi-omics datasets. The average AUROC scores on the four TCGA datasets in Table 1) highlights the strong capabilities of TabPFN-Wide. While tree-based methods exhibit stable performance, our model achieves superior results. TabPFNv2 and TabICL exhibit inferior performance consistent with the fact that they were not trained for such extreme feature counts. RealMLP-TD, trained on each

Dataset		LGG	OV	BRCA	SARC
#features		60,664	60,443	26,577	26,577
	1.5k	**0.989** ± 0.010	**0.986** ± 0.006	0.978 ± 0.002	**0.954** ± 0.005
TabPFN-Wide	5k	0.987 ± 0.008	0.985 ± 0.006	**0.984** ± 0.002	0.950 ± 0.007
	8k	**0.989** ± 0.009	0.983 ± 0.006	0.983 ± 0.000	0.953 ± 0.003
TabPFNv2		0.875 ± 0.010	0.899 ± 0.005	0.884 ± 0.004	0.902 ± 0.010
TabICL		0.943 ± 0.010	0.718 ± 0.011	0.943 ± 0.004	0.863 ± 0.019
R. Forest		**0.989** ± 0.007	0.968 ± 0.003	0.982 ± 0.003	0.942 ± 0.017
XGBoost		0.985 ± 0.008	0.971 ± 0.006	0.981 ± 0.002	0.929 ± 0.018
RealMLP-TD		0.987 ± 0.009	0.982 ± 0.005	0.981 ± 0.004	0.952 ± 0.016

Table 1: Average AUROC (±SEM) scores on 4 real-world multi-omics datasets (higher is better). We compare TabPFN-Wide, using up to 8k features for continued pre-training to TabPFNv2 and other baselines. We boldface the best values in each column.

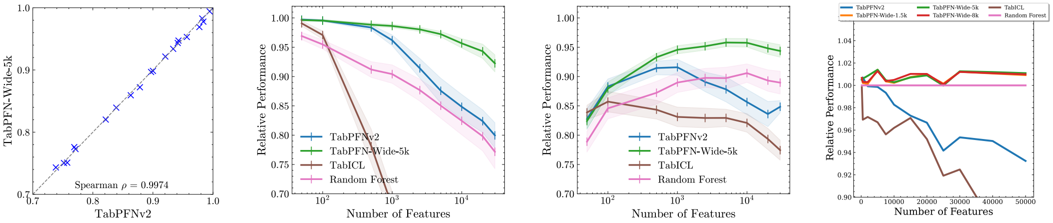
dataset separately, yields comparable although slightly inferior AUROC results to TabPFN-Wide demonstrating that it also effectively handles HDLSS data.

To enable a systematic comparison of the models across a fixed set of feature counts, we applied feature reduction. Figure 2d shows the strong relative performance for all TabPFN-Wide variants compared to a random forest. While all models perform similar with heavily reduced feature sets, the performance of TabPFN and TabICL drastically declines for higher numbers of features, while TabPFN-Wide's performance stays robust suggesting that it captures the correct signal. Notably, TabPFN-Wide exhibits competitive performance even with feature counts far exceeding those seen during continued pre-training.

Interestingly, increasing the maximum width of synthetic datasets used during continued pre-training from $1{,}500$ to $8{,}000$ exerts only a minor influence on cancer subtype classification performance, hence, further research on the optimal setting is needed.

Results on Standard Benchmarks and Widened Adaptations. Next, we compare performance on standard benchmark tasks. Figure 2a compares TabPFN-Wide to TabPFNv2 on TabArena datasets, showing that our continued pre-pretraining impacts performance negligibly, with TabPFN-Wide achieving results on par with TabPFNv2. We also generated a widened version of 13 OpenML [Bischl et al., 2025] and TabArena datasets using Algorithm 1 (see Section A.7 for details). Specifically, we explore: (a) *needle-in-a-haystack*, where we add noise features ($p = 0$, with the original features included) and (b) *feature smearing*, where the signal is distributed, i.e. "smeared", across many features ($p \in \{0.02, 0.25, 0.5\}$, without original features).

For setting (a), Figure 2b shows that our model (green line) is nearly unaffected by noisy features, resulting in only a slight performance decrease relative to TabPFNv2's performance on the original datasets. This highlights that TabPFN-Wide can pinpoint relevant features making up as little as 0.03% of all input features, i.e., the needle in the haystack. Figure 2c shows results for setting (b) where TabPFN-Wide performs again best, reaching on average about 95% of TabPFNv2's performance on the original datasets.



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

Figure 2: (a) AUROC for TabPFN-Wide-5k vs TabPFNv2 on 21 TabArena classification tasks. (b-c) Average AUROC (relative to TabPFNv2 evaluated on the original dataset) on a set of 13 widened datasets: (b) *needle-in-a-haystack* and (c) *features-smearing* (see text for further details). (d) Models' average relative performance compared to random forest (pink) for up to 4 multi-omics datasets.

5 Conclusion

We introduce TabPFN-Wide, developed by continuing pre-training of TabPFNv2. To the best of our knowledge, it is the first tabular foundation model that handles HDLSS data without feature reduction and is the first application of continued pre-training to extend tabular foundation model capabilities. It achieves state-of-the-art performance on real-world and synthetic HDLSS data while simultaneously maintaining performance on small datasets.

Since our model is currently based solely on TabPFNv2, our approach seeks further validation from continuing pre-training of the regressor model. The prior setup is strongly inspired by the biomedical data, raising the question of whether a more diverse or sophisticated HDLSS prior could further improve performance. Overall, we show that continued pre-training has the potential to extend the capabilities of pre-trained models, like TabPFNv2, paving the way for resource-efficient generation of "patched" model versions for other dataset characteristics.

Acknowledgments

Katharina Eggensperger and Nico Pfeifer acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

References

P. Bady, S. Kurscheid, M. Delorenzi, T. Gorlia, M. J. van den Bent, K. Hoang-Xuan, É. Vauléon, A. Gijtenbeek, R. Enting, B. Thiessen, O. Chinot, F. Dhermain, A. A. Brandes, J. C. Reijneveld, C. Marosi, M. J. B. Taphoorn, W. Wick, A. von Deimling, P. French, R. Stupp, B. G. Baumert, and M. E. Hegi. The DNA methylome of DDR genes and benefit from RT or TMZ in IDH mutant low-grade glioma treated in EORTC 22033. *Acta Neuropathol.*, 135(4):601–615, Apr 2018.

D. Bell, A. Berchuck, M. Birrer, J. Chien, D. W. Cramer, F. Dao, R. Dhir, et al. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, Jun 2011. URL `https://doi.org/10.1038/nature10166`.

B. Bischl, G. Casalicchio, T. Das, M. Feurer, S. Fischer, P. Gijsbers, S. Mukherjee, A. C. Müller, L. Németh, L. Oala, L. Purucker, S. Ravi, J. N. van Rijn, P. Singh, J. Vanschoren, J. van der Velde, and M. Wever. OpenML: Insights from 10 years and more than a thousand papers. *Patterns*, 6(7), Jul 2025. URL `https://doi.org/10.1016/j.patter.2025.101317`.

A. K. Bosserhoff, M. Moser, R. Hein, M. Landthaler, and R. Buettner. In situ expression patterns of melanoma-inhibiting activity (MIA) in melanomas and breast cancers. *J. Pathol.*, 187(4):446–454, Mar 1999.

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, and H. Lin, editors, *Proceedings of the 33rd International Conference on Advances in Neural Information Processing Systems (NeurIPS'20)*, pages 1877–1901. Curran Associates, 2020.

T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In B. Krishnapuram, M. Shah, A. Smola, C. Aggarwal, D. Shen, and R. Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, pages 785–794. ACM Press, 2016.

W.-C. Chen, C.-Y. Wang, Y.-H. Hung, T.-Y. Weng, M.-C. Yen, and M.-D. Lai. Systematic analysis of gene expression alterations and clinical outcomes for long-chain acyl-coenzyme a synthetase family in cancer. *PLoS One*, 11(5), May 2016.

Y. Cheng, Q. Li, G. Sun, T. Li, Y. Zou, H. Ye, K. Wang, J. Shi, and P. Wang. Serum anti-cfl1, anti-ezr, and anti-cypa autoantibody as diagnostic markers in ovarian cancer. *Scientific Reports*, 14(1), Apr 2024. URL `https://doi.org/10.1038/s41598-024-60544-2`.

R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8(1):37–49, Jan 2008. URL `https://doi.org/10.1038/nrc2294`.

R. M. de Voer, M.-M. Hahn, A. R. Mensenkamp, A. Hoischen, C. Gilissen, A. Henkes, L. Spruijt, W. A. van Zelst-Stams, C. Marleen Kets, E. T. Verwiel, I. D. Nagtegaal, H. K. Schackert, A. G. van Kessel, N. Hoogerbrugge, M. J. L. Ligtenberg, and R. P. Kuiper. Deleterious germline blm mutations and the risk for early-onset colorectal cancer. *Scientific Reports*, 5(1), Sep 2015. URL `https://doi.org/10.1038/srep14060`.

R. Dutta, P. Guruvaiah, K. K. Reddi, S. Bugide, D. S. Reddy Bandi, Y. J. K. Edwards, K. Singh, and R. Gupta. UBE2T promotes breast cancer tumor growth by suppressing DNA replication stress. *NAR Cancer*, 4(4), Dec 2022.

N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li, and A. Smola. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv:2003.06505 [stat.ML]*, 2020.

N. Erickson, L. Purucker, A. Tschalzev, D. Holzmüller, P. M. Desai, D. Salinas, and F. Hutter. Tabarena: A living benchmark for machine learning on tabular data. *arXiv preprint arXiv:2506.16791 [cs.LG]*, 2025. URL `https://arxiv.org/abs/2506.16791`.

B. Feuer, R. Schirrmeister, V. Cherepanova, C. Hegde, F. Hutter, M. Goldblum, N. Cohen, and C. C. White. Tunetables: Context optimization for scalable prior-data fitted networks. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Proceedings of the 37th International Conference on Advances in Neural Information Processing Systems (NeurIPS'24)*. Curran Associates, 2024.

R. M. French. Catastrophic interference in connectionist networks: can it be predicted, can it be prevented? In *Proceedings of the 7th International Conference on Neural Information Processing Systems*, NIPS'93, page 1176–1177, 1993.

A. Garg, M. Ali, N. Hollmann, L. Purucker, S. Müller, and F. Hutter. Real-tabpfn: Improving tabular foundation models via continued pre-training with real-world data. *arXiv preprint arXiv:2507.03971 [cs.LG]*, 2025. URL `https://arxiv.org/abs/2507.03971`.

P. Gijsbers, M. Bueno, S. Coors, E. LeDell, S. Poirier, J. Thomas, B. Bischl, and J. Vanschoren. Amlb: an automl benchmark. *Journal of Machine Learning Research*, 25(101):1–65, 2024.

X. Guo, V. Y. Jo, A. M. Mills, S. X. Zhu, C.-H. Lee, I. Espinosa, M. R. Nucci, S. Varma, E. Forgó, T. Hastie, S. Anderson, K. Ganjoo, A. H. Beck, R. B. West, C. D. Fletcher, and M. van de Rijn. Clinically relevant molecular subtypes in leiomyosarcoma. *Clin. Cancer Res.*, 21(15):3501–3511, Aug 2015.

B. Han, N. Bhowmick, Y. Qu, S. Chung, A. E. Giuliano, and X. Cui. FOXC1: an emerging marker and therapeutic target for cancer. *Oncogene*, 36(28):3957–3963, Jul 2017.

N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations (ICLR'23)*. ICLR, 2023a.

N. Hollmann, S. Müller, and F. Hutter. Gpt for semi-automated data science: Introducing caafe for context-aware automated feature engineering. *arXiv:2305.03403 [cs.AI]*, 2023b.

N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister, and F. Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637 (8045):319–326, 2025.

D. Holzmüller, L. Grinsztajn, and I. Steinwart. Better by default: Strong pre-tuned MLPs and boosted trees on tabular data. *arXiv preprint arXiv:2407.04491 [cs.LG]*, 2025. URL `https://arxiv.org/abs/2407.04491`.

S. A. Jankowski, D. S. Mitchell, S. H. Smith, J. M. Trent, and P. S. Meltzer. SAS, a gene amplified in human sarcomas, encodes a new member of the transmembrane 4 superfamily of proteins. *Oncogene*, 9(4):1205–1211, Apr 1994.

Y. Jian, L. Kong, H. Xu, Y. Shi, X. Huang, W. Zhong, S. Huang, Y. Li, D. Shi, Y. Xiao, M. Yang, S. Li, X. Chen, Y. Ouyang, Y. Hu, X. Chen, L. Song, R. Ye, and W. Wei. Protein phosphatase 1 regulatory inhibitor subunit 14C promotes triple-negative breast cancer progression via sustaining inactive glycogen synthase kinase 3 beta. *Clin. Transl. Med.*, 12(1), Jan 2022.

R. Kemker, M. McClure, A. Abitino, T. L. Hayes, and C. Kanan. Measuring catastrophic forgetting in neural networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18, 2018.

U. Krishnamurti and J. F. Silverman. HER2 in breast cancer: a review and update. *Adv. Anat. Pathol.*, 21(2):100–107, Mar 2014.

P. Langfelder and S. Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1), Dec 2008. URL `https://doi.org/10.1186/1471-2105-9-559`.

Y.-J. Lee, S.-R. Ho, J. D. Graves, Y. Xiao, S. Huang, and W.-C. Lin. CGRRF1, a growth suppressor, regulates EGFR ubiquitination in breast cancer. *Breast Cancer Res.*, 21(1), Dec 2019.

H. J. Lehtonen, T. Sipponen, S. Tojkander, R. Karikoski, H. Järvinen, N. G. Laing, P. Lappalainen, L. A. Aaltonen, and S. Tuupanen. Segregation of a missense variant in enteric smooth muscle actin γ-2 with autosomal dominant familial visceral myopathy. *Gastroenterology*, 143 (6):1482–1491, 2012. URL `https://www.sciencedirect.com/science/article/pii/S0016508512013042`.

H. Li, N. Xiao, Z. Li, and Q. Wang. Expression of inorganic pyrophosphatase (PPA1) correlates with poor prognosis of epithelial ovarian cancer. *Tohoku J. Exp. Med.*, 241(2):165–173, Feb 2017.

Y. Li, J. Park, L. Piao, G. Kong, Y. Kim, K. A. Park, T. Zhang, J. Hong, G. M. Hur, J. H. Seok, S.-W. Choi, B. C. Yoo, B. A. Hemmings, D. P. Brazil, S.-H. Kim, and J. Park. PKB-mediated PHF20 phosphorylation on ser291 is required for p53 function in DNA damage. *Cell. Signal.*, 25 (1):74–84, Jan 2013.

P.-K. Lo, J. Mehrotra, A. D'Costa, M. J. Fackler, E. Garrett-Mayer, P. Argani, and S. Sukumar. Epigenetic suppression of secreted frizzled related protein 1 (SFRP1) expression in human breast cancer. *Cancer Biol Ther*, 5(3):281–286, Mar 2006.

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *The Seventh International Conference on Learning Representations (ICLR'19)*. ICLR, 2019.

R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, G. M. Mastrogianakis, J. J. Olson, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, Oct 2008. URL `https://doi.org/10.1038/nature07385`.

M. M. Oshiro, C. J. Kim, R. J. Wozniak, D. J. Junk, J. L. Muñoz-Rodríguez, J. A. Burr, M. Fitzgerald, S. C. Pawar, A. E. Cress, F. E. Domann, and B. W. Futscher. Epigenetic silencing of DSC3 is a common event in human breast cancer. *Breast Cancer Res.*, 7(5):669–680, Jun 2005.

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.

J. Qu, D. Holzmüller, G. Varoquaux, and M. Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, V. Smith, F. Berkenkamp, and T. Maharaj, editors, *Proceedings of the 42nd International Conference on Machine Learning (ICML'25)*, Proceedings of Machine Learning Research. PMLR, 2025. URL `https://openreview.net/forum?id=0VvD1PmNzM`.

N. Rappoport and R. Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, 46(20):10546–10562, Oct 2018. URL `https://doi.org/10.1093/nar/gky889`.

J. Ren, S. Zheng, L. Zhang, J. Liu, H. Cao, S. Wu, Y. Xu, and J. Sun. MAPK4 predicts poor prognosis and facilitates the proliferation and migration of glioma through the AKT/mTOR pathway. *Cancer Med*, 12(10):11624–11640, Mar 2023.

R. Sciot. MDM2 amplified sarcomas: A literature review. *Diagnostics (Basel)*, 11(3), Mar 2021.

A. Śliwa, M. Kubiczak, A. Szczerba, G. Walkowiak, E. Nowak-Markwitz, B. Burczyńska, S. Butler, R. Iles, P. Białas, and A. Jankowska. Regulation of human chorionic gonadotropin beta subunit expression in ovarian cancer. *BMC Cancer*, 19(1), Jul 2019.

Y. A. Su, M. M. Lee, C. M. Hutter, and P. S. Meltzer. Characterization of a highly conserved gene (OS4) amplified with CDK4 in human sarcomas. *Oncogene*, 15(11):1289–1294, Sep 1997. URL `https://doi.org/10.1038/sj.onc.1201294`.

N. H. Tang and T. Toda. MAPping the ndc80 loop in cancer: A possible link between Ndc80/Hec1 overproduction and cancer formation. *Bioessays*, 37(3):248–256, Mar 2015.

T. Tsunoda, M. Riku, N. Yamada, H. Tsuchiya, T. Tomita, M. Suzuki, M. Kizuki, A. Inoko, H. Ito, K. Murotani, H. Murakami, Y. Saeki, and K. Kasai. ENTREP/FAM189A2 encodes a new ITCH ubiquitin ligase activator that is downregulated in breast cancer. *EMBO Rep.*, 23(2), Feb 2022.

B. van Breugel and M. van der Schaar. Why tabular foundation models should be a research priority. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning (ICML'24)*, volume 251 of *Proceedings of Machine Learning Research*. PMLR, 2024.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Proceedings of the 31st International Conference on Advances in Neural Information Processing Systems (NeurIPS'17)*. Curran Associates, Inc., 2017.

Z. Wang, P. Yu, Y. Zou, J. Ma, H. Han, W. Wei, C. Yang, S. Zheng, S. Guo, J. Wang, L. Liu, and S. Lin. METTL1/WDR4-mediated tRNA m(7)g modification and mRNA translation control promote oncogenesis and doxorubicin resistance. *Oncogene*, 42(23):1900–1912, Apr 2023.

X. Wu, L. Han, X. Zhang, L. Li, C. Jiang, Y. Qiu, R. Huang, B. Xie, Z. Lin, J. Ren, and J. Fu. Alteration of endocannabinoid system in human gliomas. *J Neurochem*, 120(5):842–849, Jan 2012.

Z. Yang, R. Kotoge, X. Piao, Z. Chen, L. Zhu, P. Gao, Y. Matsubara, Y. Sakurai, and J. Sun. MLOmics: Cancer multi-omics database for machine learning. *Scientific Data*, 12(1):1–9, 2025.

H.-J. Ye, S.-Y. Liu, and W.-L. Chao. A closer look at TabPFN v2: Understanding its strengths and extending its capabilities. *arXiv preprint arXiv:2502.17361 [cs.LG]*, 2025. URL `https://arxiv.org/abs/2502.17361`.

Y. Zeng, T. Dinh, W. Kang, and A. C. Mueller. Tabflex: Scaling tabular learning to millions with linear attention. *arXiv preprint arXiv:2506.05584 [cs.LG]*, 2025. URL `https://arxiv.org/abs/2506.05584`.

Q. Zhang, Z. Wang, X. Zeng, Y. Ding, and C. Wang. Evaluation of tumorous LCP1 and ADPGK as predictive biomarker for immune-related adverse events in bone and soft tissue sarcomas treated with anti-PD-1 and anti-PD-L1 antibodies. *BMC Cancer*, 25(1), Apr 2025.

D. Zhou, L. Zhang, Q. Lin, W. Ren, and G. Xu. Data on the association of CMPK1 with clinicopathological features and biological effect in human epithelial ovarian cancer. *Data Brief*, 13:77–84, Aug 2017.

LLM Usage

Large Language Models (LLMs) were used to support the paper writing process. We used OpenAI's ChatGPT-4 and -5 to polish writing, increase conciseness of sentences, and retrieve recommendations for rewriting to increase readability and the flow of the paper. We did not use LLMs to generate any content nor did we use it for interpretation / analyses of the results. All outputs of the LLMs were thoroughly reviewed and checked before including them into the paper to guarantee that the meaning and intent stayed unaffected.

A Appendix

A.1 Data Overview

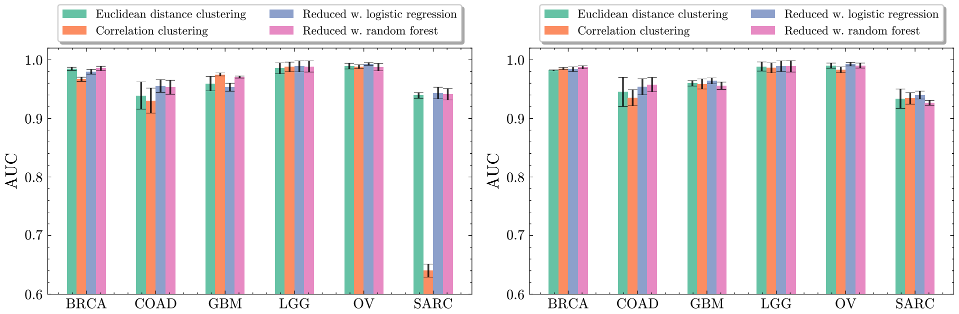
Table 2 gives an overview of the number of samples and features of the used datasets. Furthermore, it shows which molecular measurements are available for which dataset. Datasets provided by Yang et al. [2025] (COAD, LGG, OV) have 4 different omics: mRNA gene expression data (mRNA), copy number variation data (CNV), methylation data (Methylation) and micro RNA data (miRNA). MRNA, CNV, and methylation features are measurements corresponding to human genes. For our usage, we concatenated all different omics resulting in up to 60,000 features. Datasets provided by Rappoport and Shamir [2018] consist of less features due to missing CNV data and lower number of features for methylation data.

	Patients	mRNA	CNV	Methylation	miRNA	All
LGG (low grade glioma)	247	14,260	21,104	24,979	321	60,664
OV (ovarian cancer)	284	14,229	21,104	24,797	313	60,443
COAD (colon adenocarcinoma)	260	17,261	19,551	19,052	375	56,239
BRCA (breast cancer)	440	20,531	N/A	5,000	1,046	26,577
SARC (sarcoma)	259	20,531	N/A	5,000	1,046	26,577
GBM (glioblastoma)	274	12,042	N/A	5,000	534	17,576

Table 2: Number of samples and features for all used datasets. Datasets used for model selection are marked in green.

A.2 Comparison of Different Feature Reduction Techniques

In preliminary experiments, we tested the performance of TabPFNv2 on our real-world HDLSS datasets reduced with different feature reduction methods. Since this is not our main priority, we focused on simple approaches offered by *sci-kit learn* [Pedregosa et al., 2011]. Although we tested both supervised (label-based) and unsupervised feature reduction methods, our preference was for the unsupervised approaches, as they better mitigate the risk of overfitting in HDLSS settings. For biomedical data, a common approach is to cluster by correlation [Langfelder and Horvath, 2008] which we compared against clustering by lowest Euclidean distance between feature vectors and reduction using the feature importance weights from fitted machine learning models. Given that Euclidean distance-based clustering frequently outperforms the correlation-based approach for our data (see Figure 3) and achieves performance comparable to supervised methods, we adopted this strategy for our analyses.



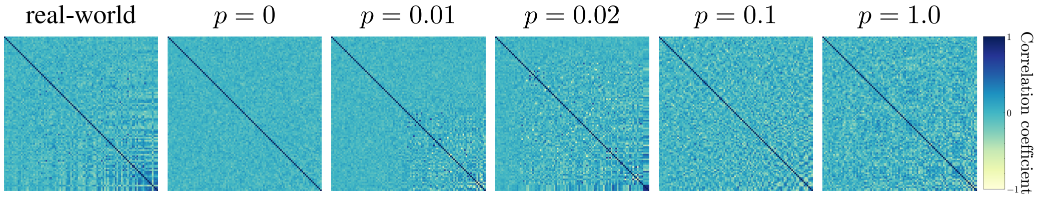
(a)

(b)

Figure 3: AUROC of TabPFNv2 evaluated on different datasets reduced to (a) 500 features and (b) 2,000 features using different techniques.

A.3 HDLSS Prior Sparsity Comparison

Our procedure can generate new features that form correlated clusters as new features depend on only a subset of the original features. The sparsity parameter p controls this structure: small values yield new features influenced by few or no originals, resulting in sparse correlation patterns, whereas large values produce new features that are mixtures of many originals, leading to dense correlation patterns. Figure 4 compares real-world HDLSS biomedical data (a) with synthetic datasets (b–f), with $p = 0.02$ showing the closest match to the real correlation structure.



(a) (b) (c) (d) (e) (f)

Figure 4: Feature correlation maps for (a) mRNA gene expression data and (b–f) synthetically generated datasets with different sparsity values p. We compute Pearson correlation for 100 randomly sampled features and sort them by average absolute correlation.

A.4 Training of TabICL with HDLSS Prior

We tried training TabICL [Qu et al., 2025] with the same training setup as for TabPFN-Wide. However, the model's training performance did not improve, suggesting that our HDLSS prior may not be effective for TabICL. Whether this arises from TabICL's architectural setup which could make it unsuitable for HDLSS data in general or whether changes to the prior / continued pre-training could mitigate this problem, remains open for future research.

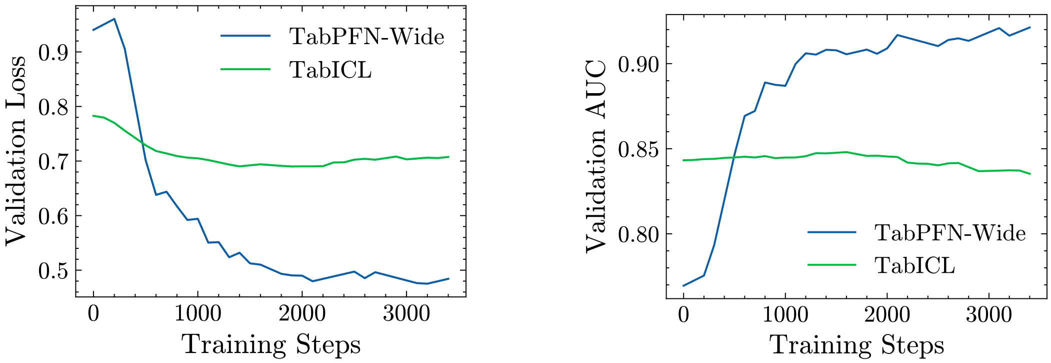
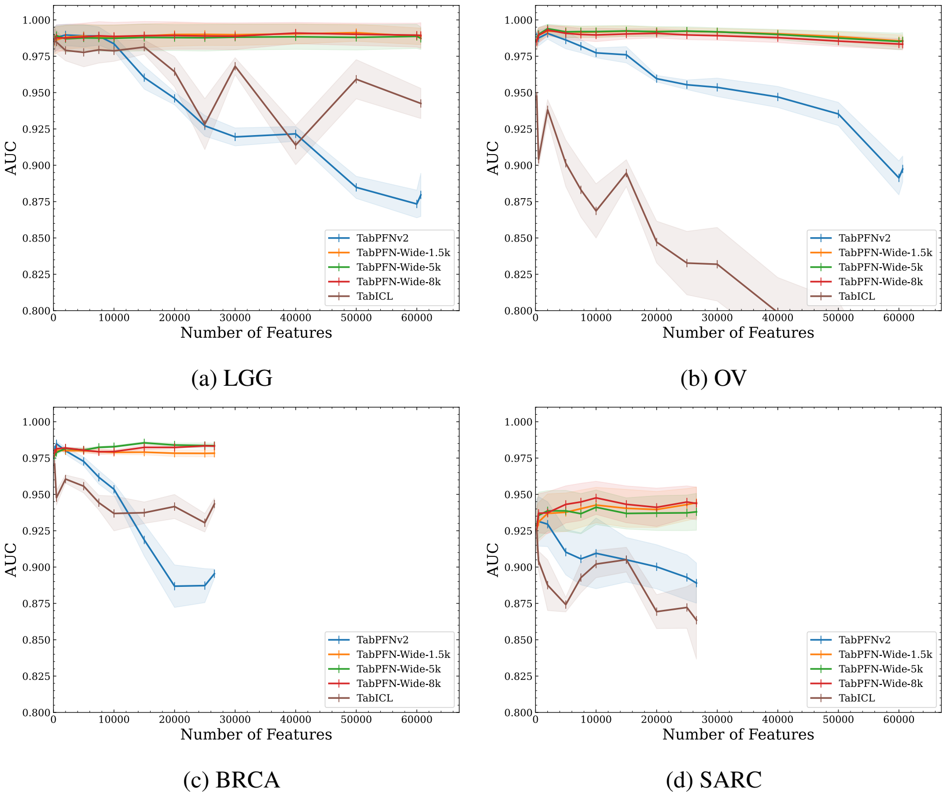


Figure 5: Development of validation loss (left) and validation AUROC (right) for TabICL vs. TabPFN-Wide when training with the same HDLSS prior.

A.5 Detailed Results for all Multi-Omics Datasets



(a) LGG

(b) OV

(c) BRCA

(d) SARC

Figure 6: Results for all datasets with feature reduction applied. The axis were chosen such that the differences in feature numbers and AUROC scores becomes comparable.

A.6 Different Metrics Analysis

We also calculated different metrics for the evaluation on our multi-omics datasets to gain a comprehensive view and address issues posed by using AUROC only.

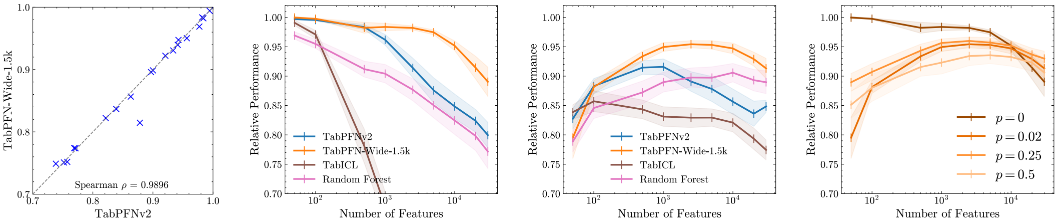
Dataset #features		LGG 60,664	OV 60,443	BRCA 26,577	SARC 26,577
	1.5k	0.980 ± 0.009	**0.965** ± 0.009	0.919 ± 0.012	**0.838** ± 0.026
TabPFN-Wide	5k	0.980 ± 0.012	**0.965** ± 0.015	0.934 ± 0.015	0.837 ± 0.032
	8k	**0.986** ± 0.010	0.960 ± 0.009	0.933 ± 0.006	0.829 ± 0.017
TabPFNv2		0.747 ± 0.014	0.795 ± 0.008	0.753 ± 0.014	0.646 ± 0.020
TabICL		0.889 ± 0.021	0.507 ± 0.006	0.817 ± 0.006	0.638 ± 0.060
R. Forest		0.983 ± 0.009	0.925 ± 0.011	0.926 ± 0.016	0.776 ± 0.025
XGBoost		0.976 ± 0.011	0.932 ± 0.012	0.928 ± 0.012	0.790 ± 0.043
RealMLP-TD		0.980 ± 0.012	0.957 ± 0.010	**0.940** ± 0.008	0.824 ± 0.042

Table 3: Average AUPRC (±SEM) scores of 4 multi-omics datasets (higher is better). We compare TabPFN-Wide, using up to 8k features for continued pre-training (second column), to TabPFNv2 and other baseline methods and boldface the best values for each column.

Dataset #features		LGG 60,664	OV 60,443	BRCA 26,577	SARC 26,577
	1.5k	0.959 ± 0.017	**0.898** ± 0.019	0.848 ± 0.009	0.772 ± 0.024
TabPFN-Wide	5k	0.972 ± 0.005	**0.898** ± 0.020	0.884 ± 0.009	0.760 ± 0.024
	8k	0.972 ± 0.010	0.887 ± 0.009	0.859 ± 0.006	0.764 ± 0.017
TabPFNv2		0.806 ± 0.006	0.679 ± 0.008	0.651 ± 0.012	0.683 ± 0.013
TabICL		0.822 ± 0.020	0.472 ± 0.014	0.768 ± 0.008	0.656 ± 0.039
R. Forest		0.956 ± 0.016	0.852 ± 0.018	0.845 ± 0.009	0.756 ± 0.029
XGBoost		**0.976** ± 0.008	0.824 ± 0.014	0.873 ± 0.012	0.761 ± 0.044
RealMLP-TD		0.964 ± 0.010	0.884 ± 0.016	**0.891** ± 0.014	**0.807** ± 0.033

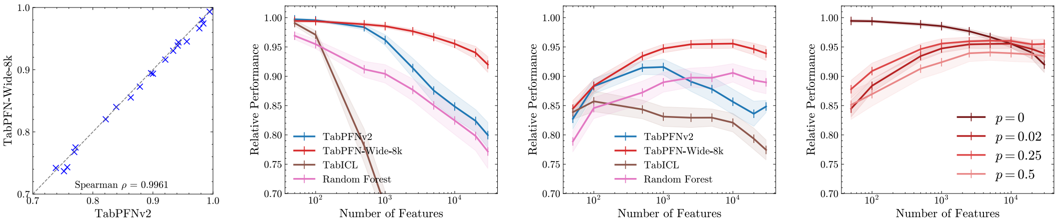
Table 4: Average accuracy (±SEM) scores of 4 multi-omics datasets (higher is better). We compare TabPFN-Wide, using up to 8k features for continued pre-training (second column), to TabPFNv2 and other baseline methods and boldface the best values for each column.

A.7 Benchmark Results for Different TabPFN-Wide Models



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

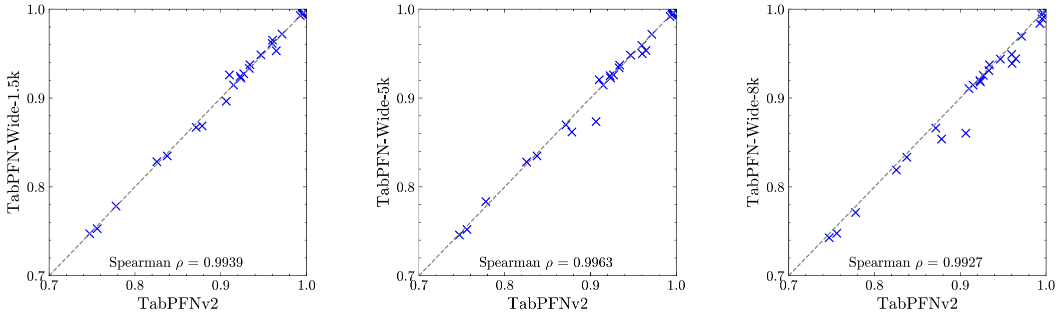
Figure 7: (a) AUROC for TabPFN-Wide-1.5k vs TabPFNv2 on 21 TabArena classification tasks with $\leq$ 10,000 samples and $\leq$ 500 features. (b-c) Average AUROC (relative to TabPFNv2 evaluated on the original dataset) on a set of 13 widened datasets: (b) *needle-in-a-haystack* and (c) *features-smearing*. (d) TabPFN-Wide-1.5k's performance for different sparsities. $p = 0$ corresponds to TabPFN-Wide-1.5k's curve in (b), and $p = 0.02$ in (c)



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

Figure 8: (a) AUROC for TabPFN-Wide-8k vs TabPFNv2 on 21 TabArena classification tasks with $\leq$ 10,000 samples and $\leq$ 500 features. (b-c) Average AUROC (relative to TabPFNv2 evaluated on the original dataset) on a set of 13 widened datasets: (b) *needle-in-a-haystack* and (c) *features-smearing*. (d) TabPFN-Wide-8k's performance for different sparsities. $p = 0$ corresponds to TabPFN-Wide-8k's curve in (b), and $p = 0.02$ in (c)

We evaluated all 3 models (TabPFN-Wide-1.5kl-5kl-8k) on the TabArena [Erickson et al., 2025] benchmark with classification datasets within TabPFNv2's sample ($\leq 10,000$) and feature (≤ 500) range. TabPFN-Wide5k has the best performance with the highest spearman correlation coefficient. TabPFN-Wide1.5k shows decent performance as well with one outlier dataset (see Figure 7). For TabPFN-Wide-8k, the performance for most datasets is slightly worse compared to TabPFNv2 showing more datasets below the diagonal compared to the other models. However, the relative and absolute performance differences are small, as seen in Figure 8. All in all, the three models maintain good performance on the TabArena benchmark, with TabPFN-Wide-5k performing best. On classification datasets within TabPFNv2's range of the AutoML benchmark [Gijsbers et al., 2024], the results are similar with TabPFN-Wide-8k decreasing most in performance (see Figure 9). Overall, TabPFN-Wide-5k shows the highest correlation coefficient with TabPFN-Wide-1.5k's coefficient being insignificantly worse, hence overall, hinting at an inverse relationship between wider datasets during training and performance on datasets within TabPFNv2's original ranges.



(a)

(b)

(c)

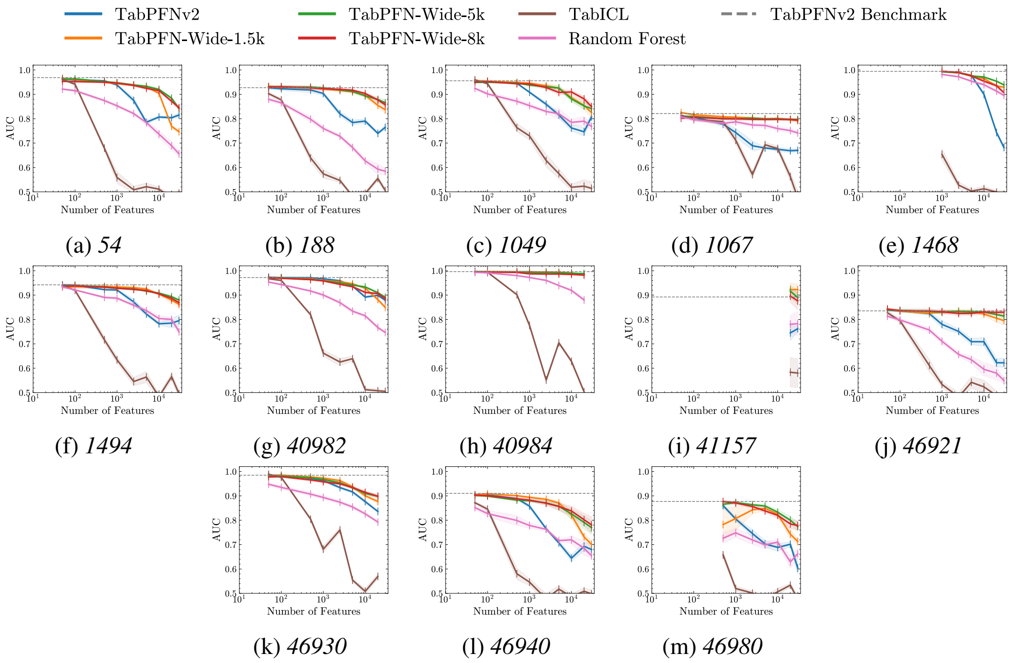
Figure 9: AUROC for TabPFN-Wide models vs TabPFNv2 on 27 AutoML benchmark classification tasks with $\leq$ 10,000 samples and $\leq$ 500 features.

For the *needle-in-a-haystack* and *feature smearing* tasks, we chose a subset of the TabArena and the AutoML benchmark. The intuition behind this selection was to evaluate TabPFN-Wide on datasets that are close to our HDLSS use case, while being synthetically generated. To include as many datasets as possible and increase the statistical significance of our analysis, we set the threshold for the maximum number of samples to 2,500. Secondly, applying Algorithm 1 entails two requirements: the features must be numerical, and their number should ideally be large to ensure that the constructed features can serve as meaningful mixtures of the originals. To increase dataset inclusion, we set this threshold to at least 8 numerical features. Since only 5 datasets meet these requirements in TabArena, we decided to include 9 classification datasets from the AutoML benchmarks as well, resulting in a total of 13 unique datasets (1 overlapping dataset).

All models exhibit high robustness against noise for the synthetically widened datasets across different number of features and choices of the sparsity parameter p. This highlights the ability of TabPFN-Wide to handle diverse types of noise / features. However, while showing competitive performance on real-world HDLSS datasets (see Section 4) TabPFN-Wide-1.5k has a stronger performance decline compared to the other two models towards high feature counts which may stem from the reduced number of features seen during training.

A.8 Detailed Widening Results for all used Datasets

Figures 10, 11, 12, and 13 show the results for every synthetically widened dataset that was selected for our widening experiments. The number of features refers to the absolute number of features in the dataset to allow for easier comparison regarding the width of a dataset. For Figure 10 the features of the original dataset were widened with different numbers of Gaussian noise features. For three datasets that showed missing values those were imputed to also allow for the evaluation of random forest and TabICL on them. Figures 11, 12, and 13 show the results for the datasets widened using Algorithm 1 with a sparsity of 0.02, 0.25, and 0.5 respectively.

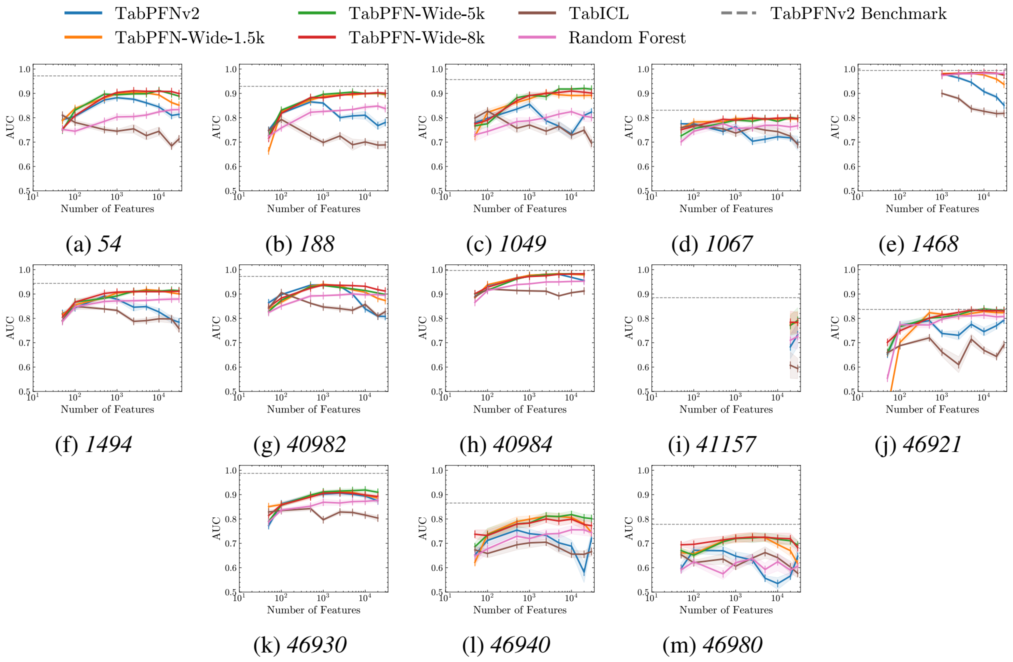


(a) *54* (b) *188* (c) *1049* (d) *1067* (e) *1468*

(f) *1494* (g) *40982* (h) *40984* (i) *41157* (j) *46921*

(k) *46930* (l) *46940* (m) *46980*

Figure 10: Detailed results for all widened OpenML datasets (*needle-in-a-haystack*). The captions indicate the corresponding OpenML dataset IDs.

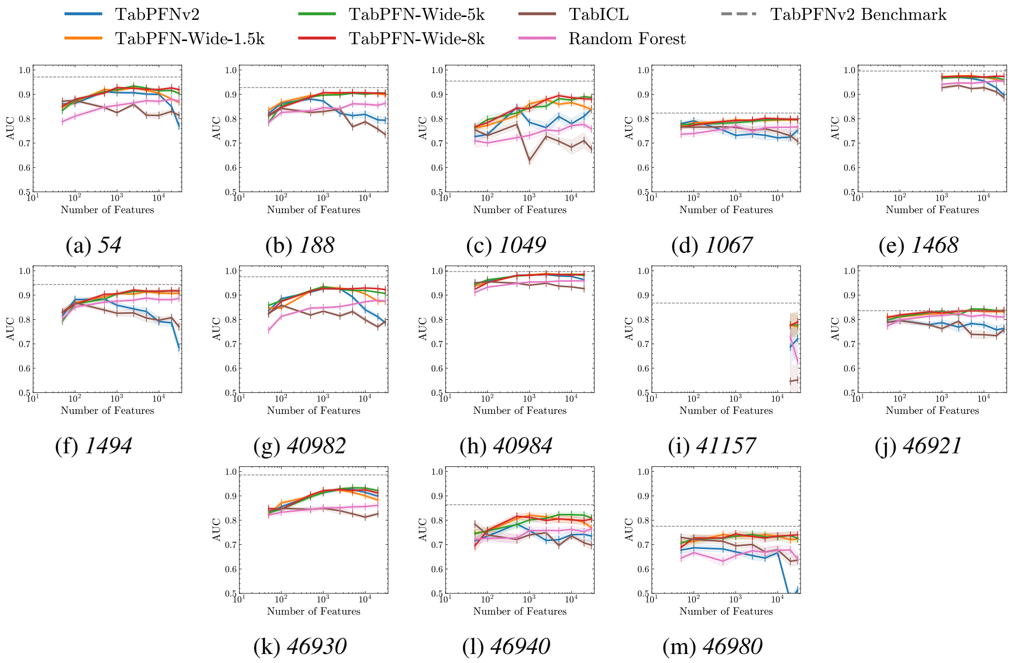


(a) *54* (b) *188* (c) *1049* (d) *1067* (e) *1468*

(f) *1494* (g) *40982* (h) *40984* (i) *41157* (j) *46921*

(k) *46930* (l) *46940* (m) *46980*

Figure 11: Detailed results for all widened OpenML datasets (*feature smearing* with $p = 0.02$). The captions indicate the corresponding OpenML dataset IDs.

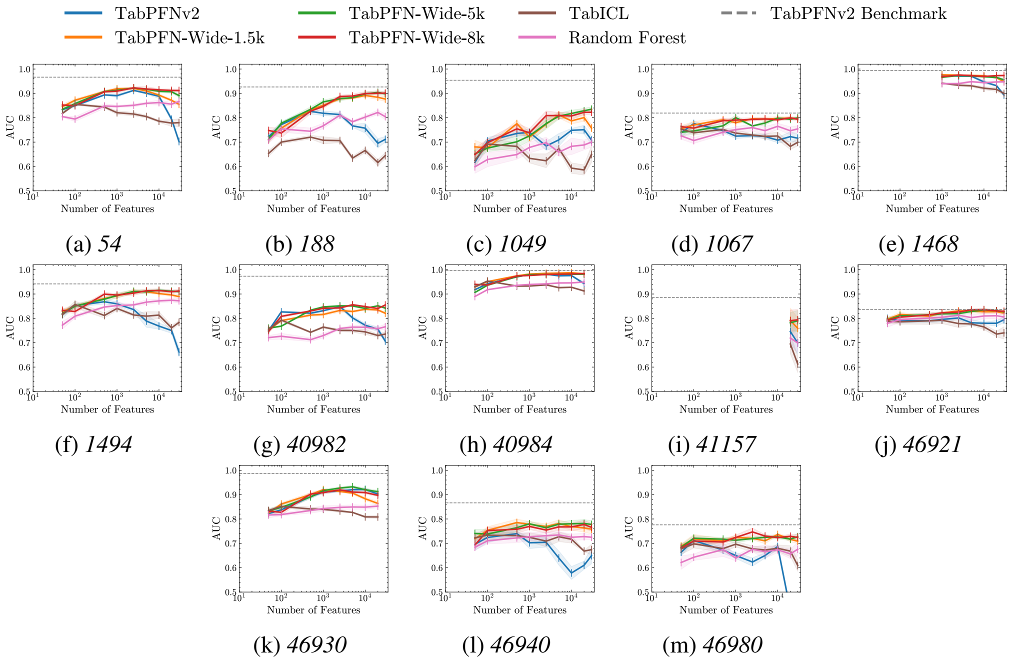


(a) *54* (b) *188* (c) *1049* (d) *1067* (e) *1468*

(f) *1494* (g) *40982* (h) *40984* (i) *41157* (j) *46921*

(k) *46930* (l) *46940* (m) *46980*

Figure 12: Detailed results for all widened OpenML datasets (*feature smearing* with $p = 0.25$). The captions indicate the corresponding OpenML dataset IDs.



(a) *54* (b) *188* (c) *1049* (d) *1067* (e) *1468*

(f) *1494* (g) *40982* (h) *40984* (i) *41157* (j) *46921*

(k) *46930* (l) *46940* (m) *46980*

Figure 13: Detailed results for all widened OpenML datasets (*feature smearing* with $p = 0.5$). The captions indicate the corresponding OpenML dataset IDs.

A.9 Feature-wise Interpretability via Attention Maps

To gain insights into TabPFNv2's inference, we analyze attention maps, focusing on attention towards the label as a proxy for feature importance. This requires that each transformer (token) column corresponds to a dataset feature. By default, TabPFNv2 groups features, adds distribution-dependent features, or may remove features impairing a token-to-feature mapping. To address this, we disabled these modifications for training as well as our biomedical datasets and interpretability analyses.

Attention maps are an intermediate step of the original dot-product attention computation [Vaswani et al., 2017] and we refer to the matrix $\boldsymbol{A}$ in Equation (1) as "attention map", with query matrix $\boldsymbol{Q}$, key matrix $\boldsymbol{K}$, value matrix $\boldsymbol{V}$, and key vector dimensionality d_{key}:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_{key}}}\right)\boldsymbol{V} = \boldsymbol{A}\boldsymbol{V}. \tag{1}$$

To interpret attention maps as an indicator of feature importance, we consider only TabPFNv2's feature-wise attention, disregarding the sample-wise attention. Since the embedded labels are appended before the forward pass, the attention value towards the label corresponds to the attention map's last row excluding the label index.

Furthermore, we average the attention maps across all samples, heads, and layers (similar to prior work by Ye et al. [2025]). We acknowledge that attention maps can vary substantially across these dimensions. However, this approach aligns with the intuition that features identified as relevant by the model across numerous samples, heads, or layers are those most indicative of importance (as we also show in our empirical results). In the following, the term "attention score" of a feature refers to its average attention to the label column.

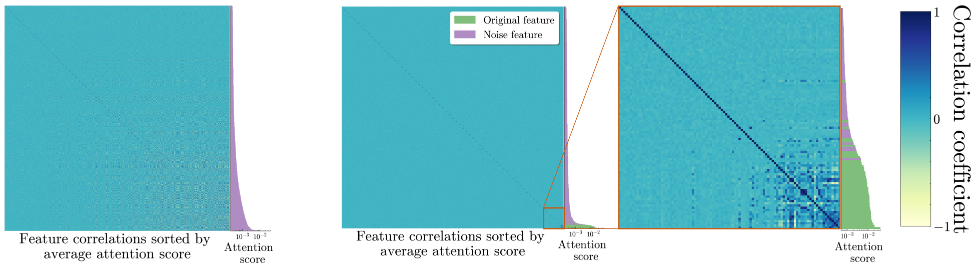
A.10 Interpretability Results

To begin our interpretability analysis, we evaluated the model on synthetically widened datasets, allowing us to assess whether attention scores reflect feature importance. Furthermore, these controlled datasets also allow us to identify, which features are expected to be predictive. We again conducted (a) *feature smearing* and (b) *needle-in-a-haystack* widening expecting our model to assign the highest scores to the original features and separate signal from noise. As described in Section A.9, we extract the attention scores for each feature during inference and average them to obtain a single value. The generated datasets contain 2,000 features and are derived from the QSAR biodegradation dataset (OpenML ID 1494). For visualization, we use correlation maps with features ordered by attention score allowing signal and noise features to be distinguished.

Features with higher attention scores are more predictive than features with lower scores. For the *feature smearing* dataset, Figure 14a shows that features with little correlation (upper left) can be distinguished from increasingly correlated features (lower right). Therefore, noisy features have low attention scores, while signal-rich features receive higher scores. The *needle-in-a-haystack* experiment further illustrates this: Figure 14b shows that the features with the highest attention scores correspond to those from the original dataset. Hence, the model not only successfully distinguished between noisy and predictive features to yield competitive performance, but this separation is also mirrored in the corresponding attention scores. These findings provide promising evidence that attention scores from TabPFN-Wide reflect feature importance and, consequently, represent a viable approach for interpretability. Results using TabPFNv2 (see Appendix A.11) show weaker separation of noise and signal, consistent with its lower performance on wide datasets.

Having evidence that attention maps yield useful insights in feature importance, we return to our real-world cancer datasets and validate the biological relevance of our model's attention scores by retrieving the features with the highest attention scores for subtype classification. Since mRNA is the most studied modality among the different omic types, we focus on the mRNA data. High correlation between genes complicates the task, since features that are presumably predictive are not necessarily causal.

TabPFN-Wide identifies important biomarkers for different cancer subtypes. We extracted the 10 genes with the highest attention scores from each dataset and examined their biological relevance according to literature (see Section A.12 for details). For breast cancer data (BRCA), all of these genes have known links to breast cancer, confirming their biological relevance and validating our



(a)

(b)

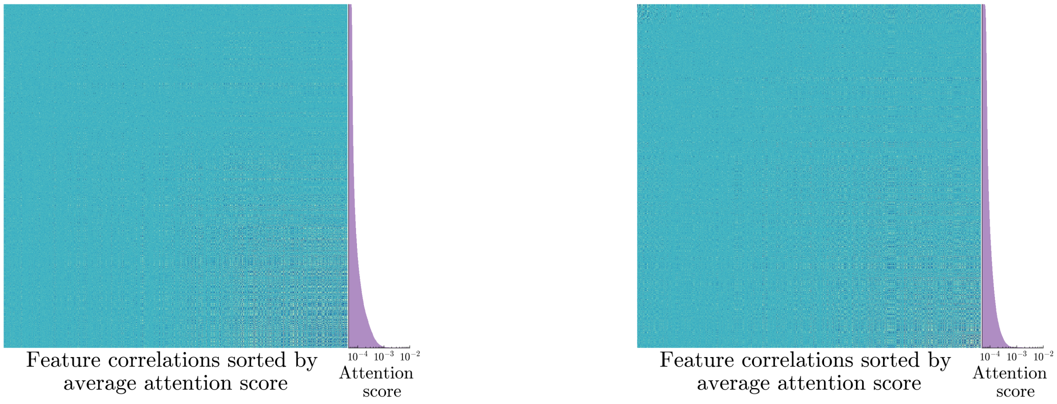
Figure 14: Correlations of 2,000 features sorted by their attention score. (a) *feature smearing* with $p = 0.02$ and $\sigma = 1$. (b) *needle-in-a-haystack*.

method. Genes such as *FOXC1*, *ERBB2*, *PPP1R14C*, and *NDC80* are directly connected to certain subtypes of breast cancer, aligning well with the subtype classification task addressed by the model. However, in other datasets fewer features could be validated by this literature review $(3/10)$. This may indicate that these cancer types are not as well studied as breast cancer, hinting at potentially undiscovered relationships, though variability in attention maps cannot be ruled out. Nevertheless, we believe these exciting results support the usefulness of attention maps as interpretability tools.

A.11 Attention Score Comparison

To compare the attention scores of TabPFNv2 and TabPFN-Wide we repeated our experiments described in Section A.10 with 10,000 features with the assumption that a reduced performance coincides with a reduced interpretability of the attention scores.

Figure 15 shows the correlations of *feature smearing* datasets. TabPFN-Wide (left) shows patterns more concentrated in the lower corner whereas TabPFNv2 pattern are far more spread with even some in the upper left corner (corresponding to lowest attention scores). This indicates that our model is better at separating noise from signal for this task.

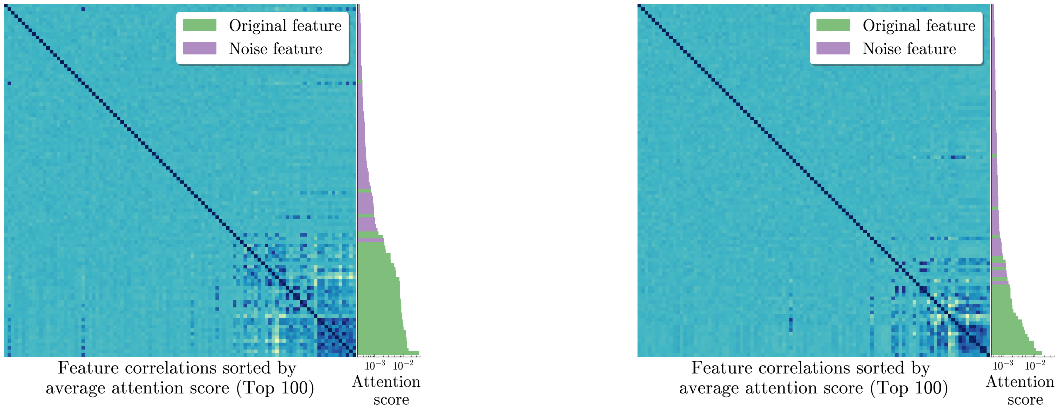


(a)

(b)

Figure 15: Comparison of correlations (TabPFN-Wide (left); TabPFNv2 (right)) between features ordered by their attention score for a *feature smearing* dataset with $p = 0.02$ and $\sigma = 1$

Figure 16 shows the correlations of the 100 features with the highest attention scores for a *needle-in-a-haystack* dataset with 10,000 features in total. Although TabPFNv2 is able to recover some of the original features, TabPFN-Wide identifies a larger number overall while also assigning higher average attention scores.



(a)

(b)

Figure 16: Comparison of correlations (TabPFN-Wide (left); TabPFNv2 (right)) between the top 100 features with the highest attention scores for a *needle-in-a-haystack* dataset with 10,000 features overall.

A.12 Genes with highest attention scores

As described in Section A.10 we analyzed the genes with the highest attention scores from our datasets with respect to literature connecting the gene with the given cancer type. We classified each gene as (i) directly associated with the specified cancer subtype, (ii) generally associated with cancer across multiple types, or (iii) having no known association with cancer. As this analysis was conducted manually, the list of citations should not be considered exhaustive. In cases where a PubMed search did not yield relevant literature, no potential associations were reported.

Dataset	**Direct Connection**	**General Connection to Cancer**	**No Known Connection**
LGG	RAD21 [Bady et al., 2018], MAPK4[Ren et al., 2023], NAPE-PLD[Wu et al., 2012]		C4B, GPN1, PPP2R3C, PRKAR1B, CWF19L2, ARIH2, PORCN
OV	CGB7[Śliwa et al., 2019], ACSL3[Chen et al., 2016], PPA1[Li et al., 2017], CFL1[Cheng et al., 2024], CGRRF1[Lee et al., 2019], CMPK1[Zhou et al., 2017]	PHF20 [Li et al., 2013],	CFD, NAXE, PDXDC1
BRCA	FOXC1 [Han et al., 2017], ERBB2 Krishnamurti and Silverman [2014], MIA [Bosserhoff et al., 1999], DSC3 [Oshiro et al., 2005], SFRP1 [Lo et al., 2006], FAM189A2 [Tsunoda et al., 2022], BLM [de Voer et al., 2015], PPP1R14C [Jian et al., 2022], NDC80 [Tang and Toda, 2015], UBE2T [Dutta et al., 2022]		
SARC	TSPAN31 [Jankowski et al., 1994], MDM2[Sciot, 2021], LMOD1[Guo et al., 2015], CTDSP2[Su et al., 1997], CDK4[Su et al., 1997], METTL1[Wang et al., 2023], ADPGK[Zhang et al., 2025], ACTG2[Lehtonen et al., 2012]		MARCH9, FAM119B

Table 5: Categorization of the top 10 features with the highest attention scores for datasets when performing subtype classification.