Transformer-based foundation Models for tabular data: handling data with arbitrary dimensions

December 5, 2025

General information

- Length of the internship: up to 6 months
- Start date: flexible in 2026
- Place: Inria Grenoble - Rhône-Alpes
- Advisors: Michael Arbel and David Salinas
- The internship can be extended to a PhD depending on the results and the candidate's interest

The internship will be held at Inria Grenoble within the Thoth team under the supervision of Michael Arbel (Research Scientist at Thoth) and David Salinas (ELLIS institute Tubingen and University of Freiburg). The student will have access to local computational resources as well as national ones (Jean-Zay) subject to approval. The student will join a dynamic, collaborative environment, surrounded by highly motivated peers and mentors who love tackling ambitious challenges together.

Project summary. Tabular and Time-series foundational models have become very popular due to their high accuracy relying solely on In-Context Learning (ICL) [1, 2]. However, one still requires training one model per "modality", e.g., one model for regression, one model for classification (although recent work showed that it is possible to reuse the TabPFN regression checkpoint for time-series [3]).

One reason for this limitation is that models use a parametric projection to obtain predictions. This parametric projection has intrinsic limitations, forcing one to learn a model that can predict only up to the maximum dimension seen during training; it also requires padding. Two exceptions are recent works which instead propose to learn a model with an equivariant architecture, with either a non-parametric approach [4] or an equivariant architecture [5].

We propose to build on this work and to train a single foundational model able to do high-dimensional classification and regression. We will leverage previous work using non-parametric approaches to search for an efficient architecture

and aim at having a single model able to perform well on TabArena [6]. If time permits, we will look into multivariate extensions, e.g., predicting joint distributions instead of marginals as done in time-series forecasting with equivariant parametrization [7] or diffusion processes [8]. We will also look into applications in tabular and time-series predictions.

Skills Technical skills:

- Strong coding ability and enthusiasm for building & running experiments at scale.
- Solid hands-on experience in deep learning (especially transformers, in-context-learning)
- Practical experience with PyTorch or JAX.

Soft skills:

- **Strong intellectual curiosity.** Eager to read papers, explore state-of-the-art ideas, and translate them into practical advances.
- **Experimentation spirit.** Thrives in the cycle of try, observe, learn and improve, embracing setbacks as essential steps toward breakthroughs.
- **Ownership & scientific rigor.** A methodical problem-solver who can track down unexpected behaviors in models or training pipelines and design clear experiments to resolve them.

Applications timeline. Applications must be sent to:

Michael Arbel (michael.arbel@inria.fr).

- Application window from 1st December to December 20th: (CV + Grades from Bachelor to Master).
- Interviews from December 22th to January 16th.

A second round might be organized if the position is not filled.

References

[1] Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *The Eleventh International Conference on Learning Representations*, 2023.

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

[3] Shi Bin Hoo, Samuel Müller, David Salinas, and Frank Hutter. From tables to time: How tabpfn-v2 outperforms specialized time series forecasting models, 2025.

[4] Mykhailo Koshil, Matthias Feurer, and Katharina Eggensperger. In-context learning of soft nearest neighbor classifiers for intelligible tabular machine learning. In *The 4th Table Representation Learning Workshop at ACL 2025*, 2025.

[5] Michael Arbel, David Salinas, and Frank Hutter. EquitabPFN: A target-permutation equivariant prior fitted network. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[6] Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Mutalik Desai, David Salinas, and Frank Hutter. Tabarena: A living benchmark for machine learning on tabular data. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[7] David Salinas, Michael Bohlke-Schneider, Laurent Callot, Roberto Medico, and Jan Gasthaus. High-dimensional multivariate forecasting with low-rank gaussian copula processes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[8] Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8857–8868. PMLR, 18–24 Jul 2021.