Evaluating TabPFN V3 for Multi-Label Toxicity Prediction: a Comparative Study on Tox21 and SIDER Datasets

Henrique Abreu Piccolo[1] **, Felipe Kenji Nakano**[2] **, Ricardo Cerri**[1]

[1] Institute of Mathematics and Computer Sciences, University of São Paulo
[2] State University of Londrina, Department of Computer Science

hpiccolo@gmail.com, kenji@uel.br, cerri@icmc.usp.br

Abstract. *Predicting molecular toxicity and side effects is a cornerstone of modern drug discovery, aimed at reducing the costs and ethical concerns associated with animal testing. This task is inherently multi-label, as a single chemical compound can interact with multiple biological targets and simultaneously trigger diverse toxicological pathways. However, tabular chemical datasets often face challenges such as extreme class imbalance and high sparsity. In such scenarios, TabPFN, a Transformer-based Prior-Data Fitted Network, emerges as a promising alternative, with studies showing its ability to generalize across a range of class imbalance ratios and its effectiveness on sparse and noisy data, often surpassing tuned XGBoost. However, none of the existing studies have investigated whether TabPFN maintains its performance in multi-label scenarios. Thus, this study investigates the performance of TabPFN V3 in predicting multi-label molecular toxicity and side effects across two benchmark datasets: Tox21 and SIDER. We represented molecular structures using 512-bit Morgan Fingerprints and investigated TabPFN V3 under two main multi-label strategies: Binary Relevance and Classifier Chains. Results on the Tox21 and SIDER datasets demonstrated that TabPFN V3 surpasses established frameworks XGBoost and Random Forest in PR-AUC scores across the evaluated multi-label spaces. However, under extreme class imbalance and positive label sparsity, these tree-based ensembles still retain a competitive advantage, yielding higher F1 scores compared to TabPFN V3. Our findings suggest that TabPFN V3 serves as a useful tool for chemical virtual screening, providing a tuning-free alternative predictive model that eliminates the operational overhead of manual hyperparameter optimization.*

1. Introduction

The early identification of toxicological profiles and adverse side effects is a bottleneck in drug discovery [Huang et al. 2016]. Methods, such as *in vivo* animal testing are ethically sensitive, expensive, and time-consuming [Raies and Bajic 2018]. Thus, computational models have become indispensable tools for virtual screening, enabling the rapid prioritization of safe molecular candidates [Huang et al. 2016, Wu et al. 2018].

This task is inherently a multi-label classification problem, since a single chemical compound can interact with a wide array of proteins and enzymes, which triggers multiple toxicological pathways or side effects simultaneously [Kuhn et al. 2016, Garcia de Lomana et al. 2021]. Modeling these interactions is challenging because chemical datasets are often imbalanced and highly sparse [Wu et al. 2018, Kuhn et al. 2016].

To tackle these problems, consolidated tree-based ensembles – such as Random Forest and Gradient Boosting variants like XGBoost – are widely adopted and remain benchmarks for tabular classification tasks [Breiman 2001, Chen and Guestrin 2016, Grinsztajn et al. 2022]. Other classical approaches, such as Support Vector Machines, are frequently explored; however, they scale poorly with large sample sizes and are highly sensitive to kernel configurations [Pedregosa et al. 2011]. Concurrently, despite the transformative impact of Deep Learning architectures, their advantages are less pronounced when applied to a tabular dataset [Grinsztajn et al. 2022]. In tabular domains, Multilayer Perceptrons (MLPs) often struggle to extract meaningful hierarchical representations from heterogeneous data without extensive architecture customization, frequently underperforming when compared to easy-to-tune tree ensembles [Grinsztajn et al. 2022].

Recently, the emergence of TabPFN has introduced a novel perspective on tabular learning by adapting a Transformer-based Prior-Data-Fitted Network to this domain [Hollmann et al. 2022]. By leveraging In-Context Learning to approximate the posterior predictive distribution in a single forward pass, TabPFN internalizes the optimization process and completely eliminates hyperparameter tuning.

In this work, we evaluate TabPFN against XGBoost and Random Forest on the Tox21 and SIDER multi-label datasets from MoleculeNet [Wu et al. 2018]. More specifically, we compare Binary Relevance (BR) and Classifier Chains (CC) frameworks [Raies and Bajic 2018], using these models as base learners.

The remainder of this paper is organized as follows. Section 2 reviews related work on molecular toxicity prediction and multi-label frameworks. Section 3 details the Tox21 and SIDER datasets, together with the pre-processing and molecular representation techniques. Section 4 describes the baseline ensembles, the TabPFN architecture, and the multi-label strategies. Section 5 discusses the experimental results, comparing the performance and computational efficiency of the models. Finally, Section 6 presents conclusions and describes directions for future work.

2. Related work

Computational predictive modeling in toxicology has increasingly shifted from traditional single-endpoint architectures to multi-label classification. This shift is motivated by the need to replicate biological systems, in which a single chemical compound can simultaneously trigger multiple toxicological pathways or interact with diverse target proteins. The relevance of multi-label learning in this domain has been extensively demonstrated. For instance, Raies and Bajic [Raies and Bajic 2018] provided a comprehensive benchmark of multi-label classification methods applied to chemical toxicity data, investigating approximately 19,000 combinations of multi-label configurations and demonstrating that modeling joint label dependencies significantly improves predictive performance compared with independent single-label tasks. Similarly, investigations in *in vitro* toxicology, such as those by Garcia de Lomana et al. [Garcia de Lomana et al. 2021], have successfully leveraged multi-label structures to predict complex endpoints, such as drug-induced liver injury targets. Studies such as [Yap 2020] reinforce the advantage of multi-label models in handling the high dimensionality and partial-label availability typical of environmental and pharmacological datasets.

While benchmarks provided by repositories like MoleculeNet have standardized

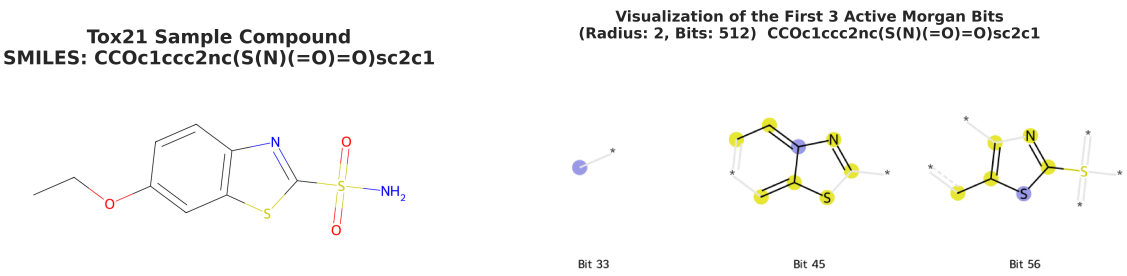
the evaluation of machine learning algorithms in cheminformatics [Wu et al. 2018], comparisons between traditional tree-based ensembles and deep learning models predominantly focus on established structural network transformations or standard neural classifiers. Consequently, they do not yet include recent, emerging tabular architectures like TabPFN. We introduce a novel comparative evaluation to bridge this gap.

3. Datasets and Pre-processing

As aforementioned, our experiments were conducted in two widely used benchmark datasets from MoleculeNet: Tox21 and SIDER. Their chemical compounds are provided in the SMILES (Simplified Molecular Input Line Entry System) format, which serves as the primary basis for generating the input features. The targets for the classification task are defined by the biological or physiological outcomes associated with each molecule. Specifically, Tox21 comprises approximately 8,000 compounds, with targets spanning 12 biological pathways, including nuclear receptors and stress response pathways [Huang et al. 2016]. SIDER (Side Effect Resource) contains 1,427 drugs where the targets are categorized into 27 physiological systems, reflecting adverse drug reactions (ADRs) recorded in humans [Kuhn et al. 2016].

3.1. Molecular Representation

The SMILES notation represents molecules as linear strings of text, where letters denote atoms and symbols describe bonds and branching. Figure 1a illustrates a typical compound in this notation. However, SMILES strings cannot be input directly into most machine learning algorithms as these models require fixed-length vectors to perform mathematical operations. Thus, we employed a fingerprinting approach to transform the qualitative chemical strings into quantitative descriptors. This process deconstructs the molecule into hierarchical atom environments, where each fragment is hashed to a specific bit index [Rogers and Hahn 2010] as shown in Figure 1b. Consequently, we adopted the **ECFP4** configuration for representing the fingerprints [Rogers and Hahn 2010]. Although alternative representations exist—such as 2D/3D graph structures, continuous learned embeddings, and physicochemical descriptors—we prioritized 512-bit ECFP4 fingerprints because they constitute the standardized, tabular baseline validated by MoleculeNet [Wu et al. 2018, Rogers and Hahn 2010].



(a) Tox21 sample compound.

(b) First 3 active Morgan bits (Radius 2).

Figure 1. Structural basis and fragment extraction for 512-bit Morgan Fingerprints.

3.2. Dataset Characterization and Imbalance Analysis

As illustrated, the dataset encompasses three distinct scenarios: classes with positive imbalance, in which side effects are reported for nearly all drugs; classes in statistical equilibrium; and categories with negative imbalance, characterized by rare adverse reactions, as can be seen in Figure 2

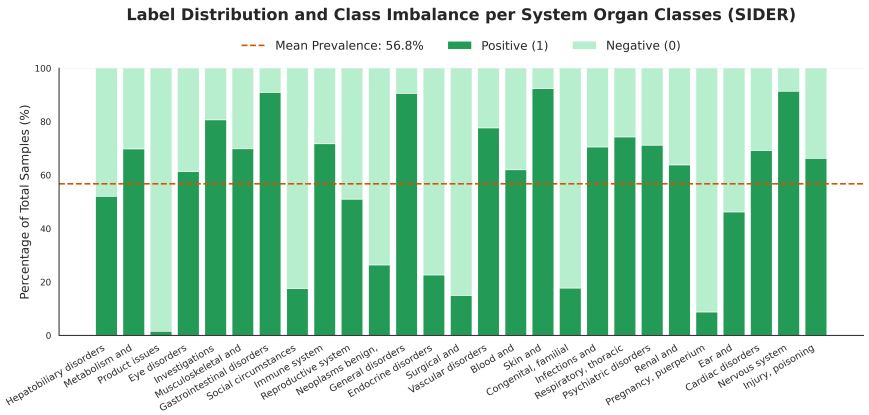


Figure 2. Distribution across the 27 system organs in SIDER. To improve visualization, x-axis labels have been truncated to the first two words.

The Tox21 dataset presents a dual challenge of imbalance and sparsity. As illustrated in Figure 3, the raw dataset comprises 7,831 compounds but contains substantial missing experimental values, as not every compound was tested for all 12 biological pathways. To enable a standardized multi-label evaluation, we filtered the data to retain only complete cases, retaining 3,079 compounds (39.3%) and removing 4,752 incomplete samples (60.7%) to generate the filtered dense subset shown in Figure 4, which is a necessary condition for the execution of certain multi-label algorithms. However, this complete-case filtering introduces a selection bias that drastically reduces the mean positive prevalence from 6.2% to 2.9% and restricts direct comparability with single-task literature evaluated on full, partially-labeled splits.

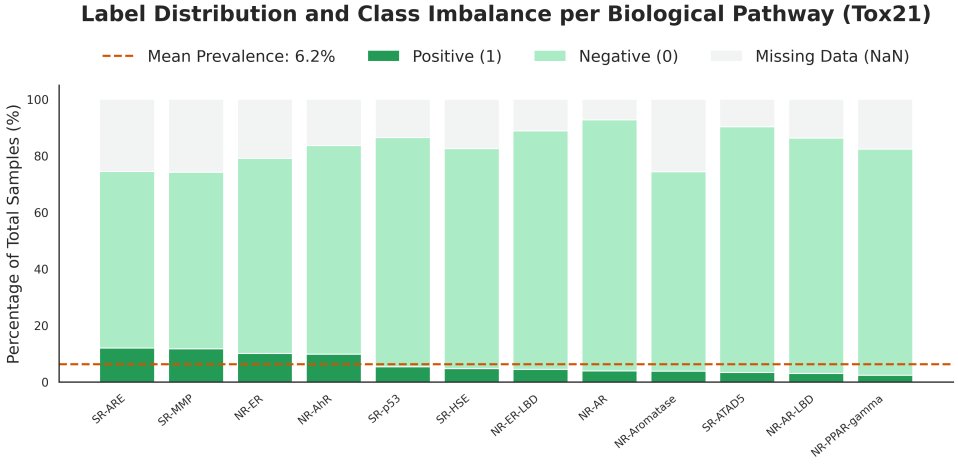


Figure 3. Distribution of biological pathways in the Tox21 dataset, including missing experimental data (NaNs).

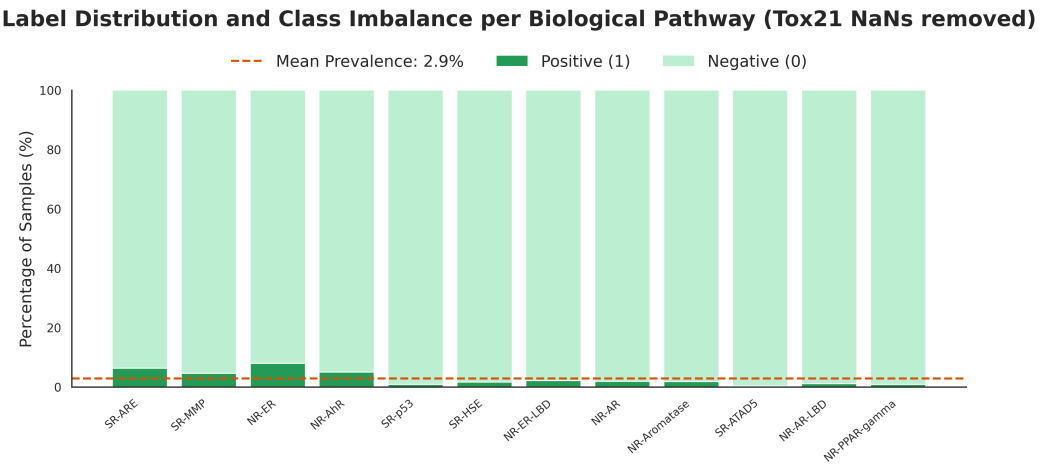


Figure 4. Class distribution in the Tox21 post-filtering.

The visual distributions observed in Figures 2, 3, and 4 provide the technical justification for our evaluation. Given that a naive classifier could achieve high accuracy by simply predicting the majority (negative) class, we prioritize Precision-Recall AUC (PR-AUC) and F1-Score as our primary metrics to ensure a robust assessment of each model.

4. Machine Learning Architectures

We selected consolidated models from the literature and contrasted them with the novel transformer-based approach TabPFN using the `tabpfn-v3-classifier-v3_default` configuration, running on an NVIDIA RTX 4060TI 8GB GPU. All experiments used 5-fold cross-validation with the multi-label Iterative Stratification [Sechidis et al. 2011] implemented in the scikit-multilearn library [Szymanski and Kajdanowicz 2019]. This stratified approach ensures that the distribution of rare toxicological pathways and adverse reactions is preserved across all training and testing partitions. All our datasets and implementations are freely available[1].

4.1. Consolidated Ensembles: RF and XGBoost

We used Random Forest (RF) and XGBoost due to their proven efficiency in handling tabular data and high-dimensional chemical spaces[Wu et al. 2018], with execution benchmarks performed on an AMD Ryzen 7 7700X CPU. The RF model was implemented using scikit-learn configured with 100 trees. The XGBoost implementation used the official xgboost Python package, with 100 boosting rounds and the log-loss evaluation metric. For both algorithms, all remaining hyperparameters were kept at their default values.

4.2. Prior-Data Fitted Network (TabPFN)

Unlike traditional models that require iterative optimization and manual hyperparameter tuning for each new dataset, TabPFN is a Transformer-based Prior-Data Fitted Network [Hollmann et al. 2022]. It was pre-trained on millions of synthetic datasets, learning to approximate the posterior predictive distribution through a single forward pass.

[1] `https://github.com/hpiccolo/eniac_tabpfnv3`

The architecture utilizes an In-Context Learning paradigm, in which the entire training set (features and labels) and the test samples are fed into the transformer simultaneously. While multiple iterations of this architecture have been developed, this study specifically leverages the latest third-generation version (TabPFN v3), using the `tabpfn-v3-classifier-v3_default` model configuration.

4.3. Multi-label Classification Strategies

We focused on two multi-label strategies, Binary Relevance and Classifier Chains [Tsoumakas et al. 2009, Read et al. 2011], as they are often used as comparison methods in toxicity studies, such as [Raies and Bajic 2018]

- **Binary Relevance (BR):** decomposes the multi-label problem into L independent binary classification tasks, where a separate model is trained for each biological pathway or organ class. This strategy is effective for maximizing the use of available data per target and serves as a baseline that ignores potential correlations between labels [Tsoumakas et al. 2009];
- **Classifier Chains (CC):** addresses the limitations of independent modeling by accounting for label correlations. Classifiers are linked in a chain, with the predictions of previous models passed as additional input features to subsequent ones. It is particularly relevant for pharmacological and toxicological data, where outcomes in different biological systems often exhibit underlying dependencies [Read et al. 2011].In this work, the chain sequence was determined using a single random ordering.

5. Results and Discussion

This section presents the empirical evaluation of TabPFN V3 against Random Forest and XGBoost across the Tox21 and SIDER benchmarks. First, we compare the performance considering all labels at once. Subsequently, we conduct a target-by-target analysis on individual pathways and organ systems to explore these performance trade-offs in detail, concluding with an assessment of computational execution runtimes.

5.1. Multi-label Analysis

In this experiment, the models were evaluated using an integrated multi-label framework implemented through the `sklearn.metrics` module. To account for the severe class imbalances, we prioritize the macro-averaged F1-score alongside the area under the Precision-Recall curve (PR-AUC) as our primary evaluation metrics.

5.1.1. Toxicology Analysis: Tox21 Dataset

Figure 5 summarizes the global performance comparing Binary Relevance (BR) and Classifier Chains (CC). Regarding PR-AUC, TabPFN V3 demonstrated global superiority, with its Binary Relevance variant achieving the highest ranking score of 0.235, followed by TabPFN V3 (CC) at 0.217. In the F1-Macro analysis, XGBoost (BR) achieved the highest performance at 0.139, while TabPFN V3 (BR) trailed significantly behind at 0.067.

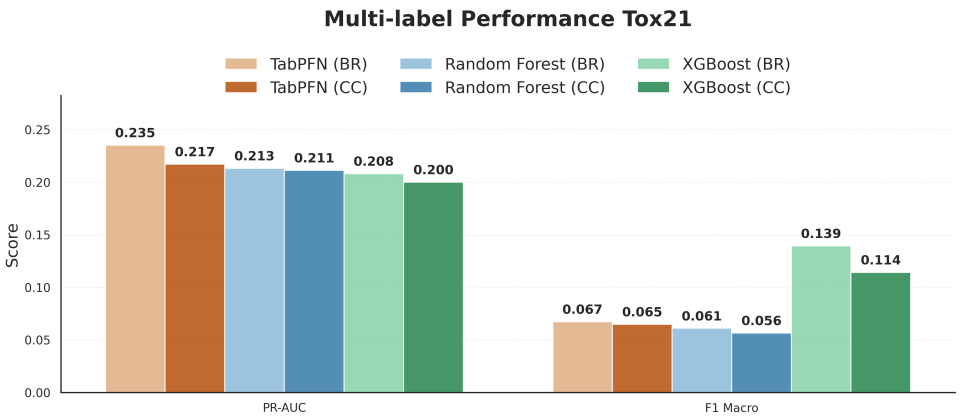


Figure 5. F1-Macro and PR-AUC on the Tox21 dataset using BR and CC.

The implementation of CC did not yield performance gains for the Tox21 dataset. For TabPFN V3, executing the sequential chain reduced the PR-AUC from 0.235 to 0.217, and slightly degraded the F1-Macro score from 0.067 to 0.065 compared to its BR counterpart. A similar downward trend was observed for Random Forest, where the CC strategy led to an F1-Macro of 0.056, underperforming relative to its BR score of 0.061. These findings suggest that a single, fixed sequential ordering may introduce error propagation throughout the chain. Consequently, a single random chain sequence might be sub-optimal, whereas an Ensemble of Classifier Chains (ECC) could mitigate this dependency by averaging multiple random orderings.

5.1.2. Pharmacological Analysis: SIDER Dataset

The evaluation on the SIDER dataset demonstrates the performance of the models in predicting adverse drug reactions (ADRs) across 27 system-organ classes. Figure 6 presents the results using the BR and CC strategies, focusing on the F1-Macro and PR-AUC metrics.

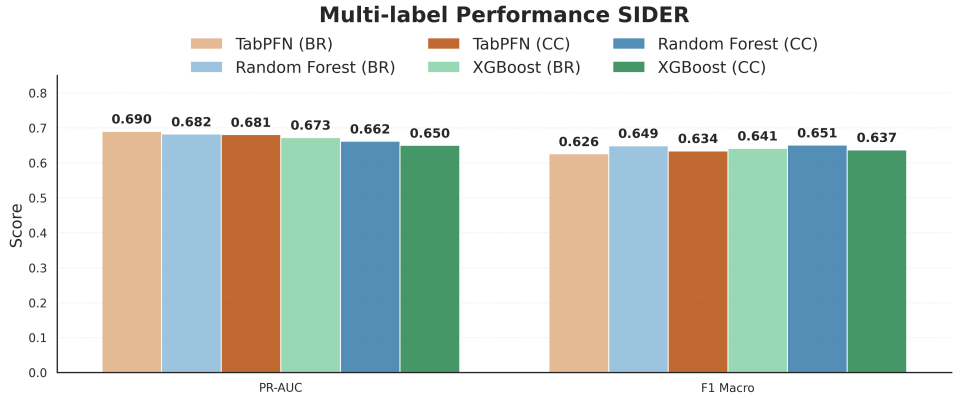


Figure 6. F1-Macro and PR-AUC on the SIDER dataset using BR and CC.

The evaluation on the SIDER dataset mirrors the overarching patterns observed in the Tox21 analysis. Regarding PR-AUC, TabPFN V3 (BR) achieved the highest overall score of 0.690, establishing a marginal lead over traditional benchmarks such as Random

Forest (BR) at 0.682 and XGBoost (BR) at 0.673. This overall dominance in ranking accuracy indicates that TabPFN V3 remains robust when generalizing across a large number of human physiological targets simultaneously. Conversely, the F1-Macro metric reflects narrow performance gap among all configurations. In this metric layer, Random Forest (CC) secured a slight advantage at 0.651, closely followed by its BR counterpart at 0.649 and XGBoost (BR) at 0.641, while TabPFN V3 (CC) achieved a competitive score of 0.634.

Overall, the multi-label evaluation across both the Tox21 and SIDER datasets confirms a recurring architectural trade-off: TabPFN consistently demonstrates global superiority in ranking accuracy via PR-AUC, yet it faces challenges in maximizing the F1-Macro metric under severe class imbalance. Furthermore, the transition from Binary Relevance (BR) to Classifier Chains (CC) did not yield the expected performance gains, often resulting in a slight degradation. It is critical to emphasize, however, that this bottleneck may stem from the stochastic nature of defining a single, static chain sequence, which can propagate early classification errors throughout the sequential target space.

5.2. Individual Pathway Analysis

As a complementary step to the global evaluations, we conducted a granular analysis of individual biological pathways and organ systems. For this target-by-target assessment, we focused exclusively on the Binary Relevance (BR) strategy, as it demonstrated superior performance over Classifier Chains (CC) in the overarching multi-label metrics. Consequently, this target-wise isolation provides a clean overview to observe how TabPFN V3 and the tree ensembles handle specific biological characteristics independently, shedding light on the components driving the macro behavior.

5.2.1. Precision Performance

We first examined the Precision metric across all targets. This auxiliary analysis helps evaluate how effectively each model identifies true positives within its toxicological and pharmacological predictions. Figures 7 and 8 present these baseline results.

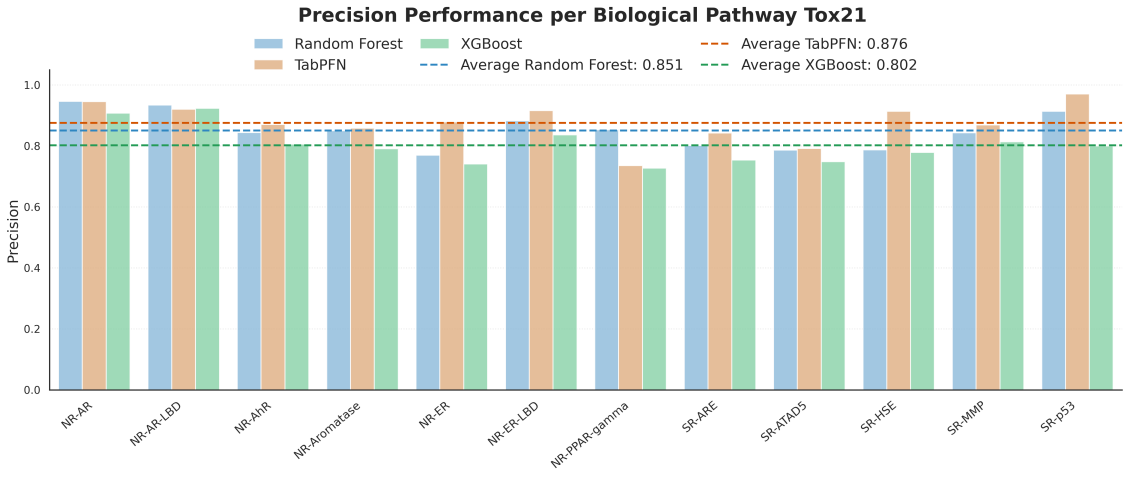


Figure 7. Precision scores across Tox21 biological pathways.

As shown in Figure 7, TabPFN V3 achieved the highest overall predictive consistency on the Tox21 dataset, leading with an average precision score of 0.876. Traditional tree-based ensembles followed, with Random Forest reaching a macro-average of 0.851, while XGBoost slightly underperformed, trailing behind with an average of 0.802.

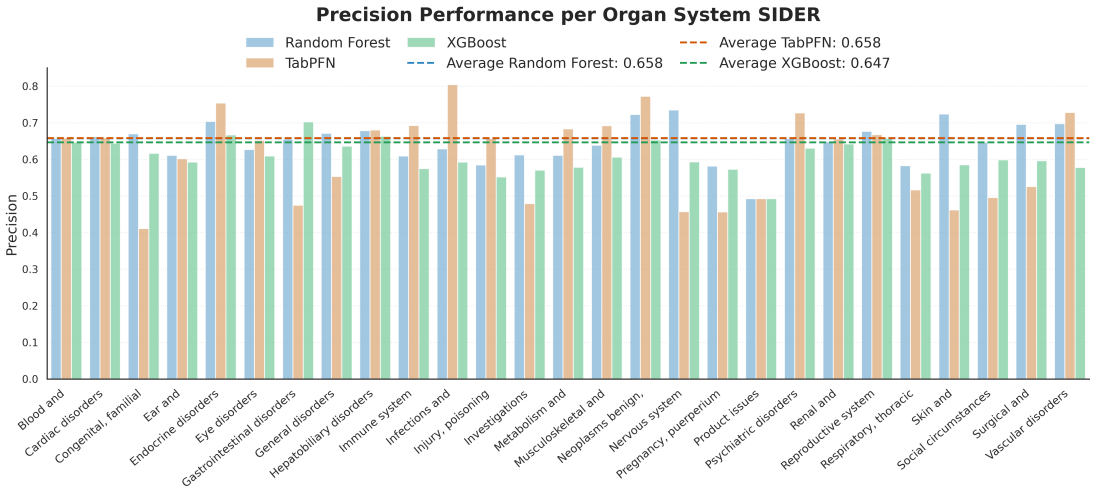


Figure 8. Precision scores across SIDER organ systems. To improve visualization, x-axis labels have been truncated to the first two words.

In the SIDER dataset (Figure 8), the predictive performance gap among the models was exceptionally narrow, resulting in a competitive tie for the leading position. Both TabPFN V3 and Random Forest achieved an identical top macro-average precision of 0.658, while XGBoost followed closely behind with an average score of 0.647.

5.2.2. F1 Performance

To provide additional support for the comparative analysis, we also evaluated the F1 score. This metric serves as an auxiliary indicator of consistency by aggregating results across all instances. Figures 9 and 10 illustrate these complementary results.

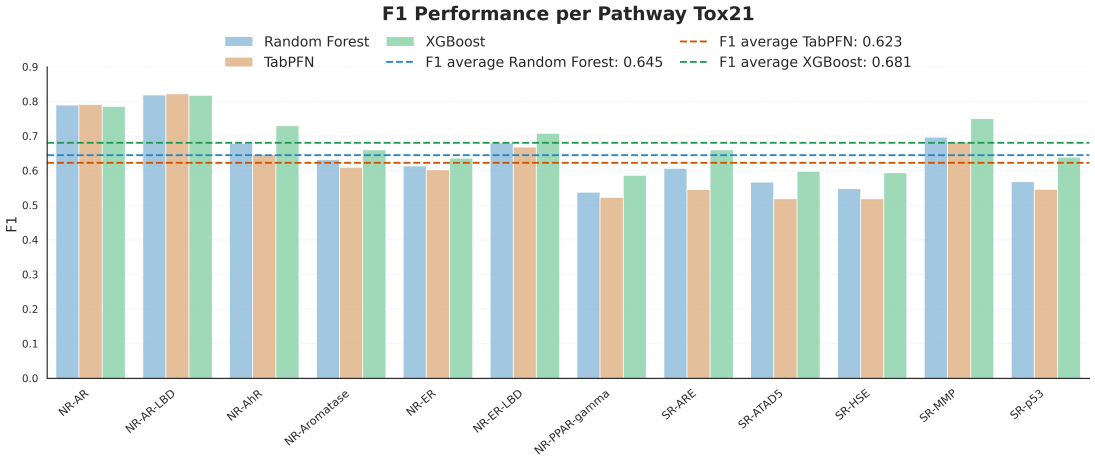


Figure 9. F1 scores across Tox21 biological pathways.

In the Tox21 dataset (Figure 9), the models exhibited a high degree of stability across the individual toxicological assays. XGBoost achieved the highest overall performance, leading with a macro-averaged F1-score of 0.681. Random Forest followed with an average score of 0.645, while TabPFN V3 delivered a slightly lower average of 0.623.

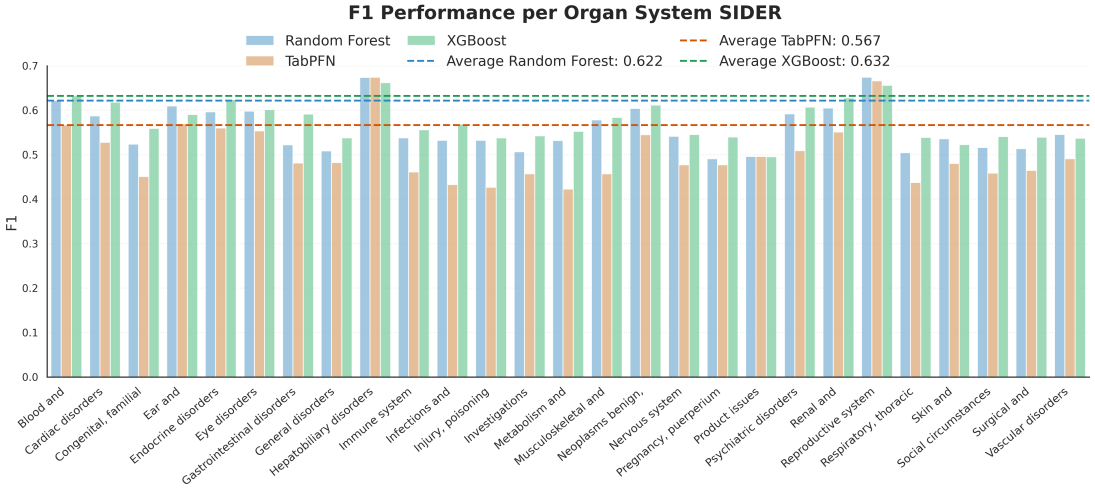


Figure 10. F1 scores scores across SIDER organ systems. To improve visualization, x-axis labels have been truncated to the first two words.

For the SIDER dataset (Figure 10), XGBoost secured the highest overall performance layer, leading with F1-score of 0.632, followed closely by Random Forest at 0.622. TabPFN V3 delivered a score of 0.567.

The individualized pathway analysis demonstrates that model superiority is not absolute. The performance of TabPFN V3 across specific biological pathways positions it as an conservative classifier optimized to eliminate false positives. Conversely, the robustness displayed by XGBoost and Random Forest in terms of F1-scores reflects their superior capacity in dealing with data imbalance, effectively capturing a broader spectrum of positive signals without discarding rare target interactions. Interestingly, in more balanced SIDER dimensions where class distribution sits closer to the mean prevalence—such as Hepatobiliary disorders and Reproductive system disorders,TabPFN V3 demonstrates competitiveness, even outperforming XGBoost in F1-score.

5.3. Computational Cost and Execution Time

Although computational efficiency was not the primary focus of this evaluation, there is a massive disparity in processing time between the architectures. All experiments were conducted on the same local workstation, but TabPFN was specifically executed on the GPU via CUDA, as its transformer attention mechanism makes CPU-only inference computationally prohibitive. As opposed to that, GPUs do not support traditional tree-based ensembles, nonetheless CPUs, in this case, are capable of executing them instantaneously. In the Tox21 dataset, for instance, XGBoost and Random Forest recorded negligible average processing times per fold execution (0.13 and 0.18 seconds, respectively). Conversely, TabPFN V3 demanded a significantly higher computational overhead, averaging 12.25 seconds per execution block, which resulted in a total processing time nearly 100 times greater than that of the ensembles.

6. Conclusion

This study evaluated the performance of TabPFN V3, a Transformer-based Prior-Data Fitted Network, for multi-label toxicological and pharmacological classification. By comparing it against established ensembles such as Random Forest and XGBoost, using Classifier Chains and Binary Relevance strategies, across the Tox21 and SIDER datasets, we observed that TabPFN V3 is a viable model for cheminformatics applications.

Our findings indicate that TabPFN V3 surpasses long-consolidated machine learning frameworks in ranking capacity. In the Tox21 dataset, the model achieved the highest multi-label PR-AUC of 0.235 under Binary Relevance (BR), outperforming Random Forest BR (0.213) and XGBoost BR (0.208), while leading several individual pathways with an average Precision of 0.876. In the SIDER dataset, TabPFN V3 BR reasserted this robust ranking performance, leading with a PR-AUC of 0.690. Conversely, tree-based ensembles retained their advantage in F1-Macro under both strategies, reaching 0.139 on Tox21 (XGBoost BR) and 0.651 on SIDER (Random Forest CC).

In the SIDER benchmark, TabPFN V3 remained competitive in F1-Macro, closely trailing tree-based ensembles. In contrast, the Tox21 dataset was associated with a sharp drop in F1-Macro across all evaluated models relative to their strong PR-AUC scores, with TabPFN V3 experiencing a more acute degradation (0.067 vs. 0.139 for XGBoost). This widespread decline is directly driven by the extreme label sparsity of Tox21 (mean prevalence of 2.9%), where models operating under the default 0.5 threshold can be heavily penalized. To mitigate this problem, future implementations could incorporate dynamic, task-specific thresholds.

Despite the promising results demonstrated in this study, it is important to acknowledge certain limitations. First, the evaluation was restricted to two specific datasets (Tox21 and SIDER); therefore, these findings should not be broadly generalized to all chemical spaces without further extensive validation. Nonetheless, this work demonstrates that TabPFN V3 serves as a highly competitive paradigm for chemical virtual screening, delivering superior PR-AUC performance without requiring manual hyperparameter tuning. Future research will explore broader baseline models, such as LightGBM, CatBoost, and Multi-Layer Perceptrons, alongside alternative molecular representations (e.g., learned graph embeddings), while scaling these networks to larger chemical libraries and optimizing inference latency.

References

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Garcia de Lomana, M. et al. (2021). Multi-label classification models for predicting drug-induced liver injury targets. *Toxicology in Vitro*, 72:105094.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? *Advances in Neural Information Processing Systems*, 35:507–523.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. (2022). TabPFN: A transformer that solves small tabular classification problems in a second. In *NeurIPS 2022 First Table Representation Workshop*.

Huang, R., Xia, M., Nguyen, D. T., et al. (2016). Tox21 challenge to predict the effects of toxic compounds on biological pathways. *Frontiers in Environmental Science*, 3:85.

Kuhn, M., Letunic, I., Jensen, L. J., and Bork, P. (2016). The sider database of drugs and side effects. *Nucleic Acids Research*, 44(D1):D1075–D1079.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Raies, S. and Bajic, V. B. (2018). In silico toxicology: comprehensive benchmarking of multi-label classification methods applied to chemical toxicity data. *WIREs Computational Molecular Science*, 8(3):e1352.

Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, 85(3):333–359.

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754.

Sechidis, K., Tsoumakas, G., and Vlahavas, I. (2011). On the stratification of multi-label data. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158, Berlin, Heidelberg. Springer-Verlag.

Szymanski, P. and Kajdanowicz, T. (2019). Scikit-multilearn: a scikit-based python environment for performing multi-label classification. *J. Mach. Learn. Res.*, 20(1):209–230.

Tsoumakas, G., Katakis, I., and Vlahavas, I. (2009). Mining multi-label data. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*. Springer, Boston, MA.

Wu, Z. et al. (2018). Moleculenet: a benchmark for deep learning in molecular design. *Chemical Science*.

Yap, X. H. (2020). Multi-label model for toxicity prediction. *Celebration of Research, Scholarship, and Creative Activities, Wright State University*, 37.