Better by Default: Strong Pre-Tuned MLPs and Boosted Trees on Tabular Data

David Holzmüller
INRIA

Léo Grinsztajn
INRIA

Ingo Steinwart
University of Stuttgart

NeurIPS 2024

Motivation: Supervised learning on tabular data

Table: A typical table (rows = samples, columns = features).

Age	Gender	Weight	Time in Hospital	...	Readmitted?
60	Male	90	5 days	...	Yes
45	Female	N/A	3 days	...	No
⋮	⋮	⋮	⋮		⋮

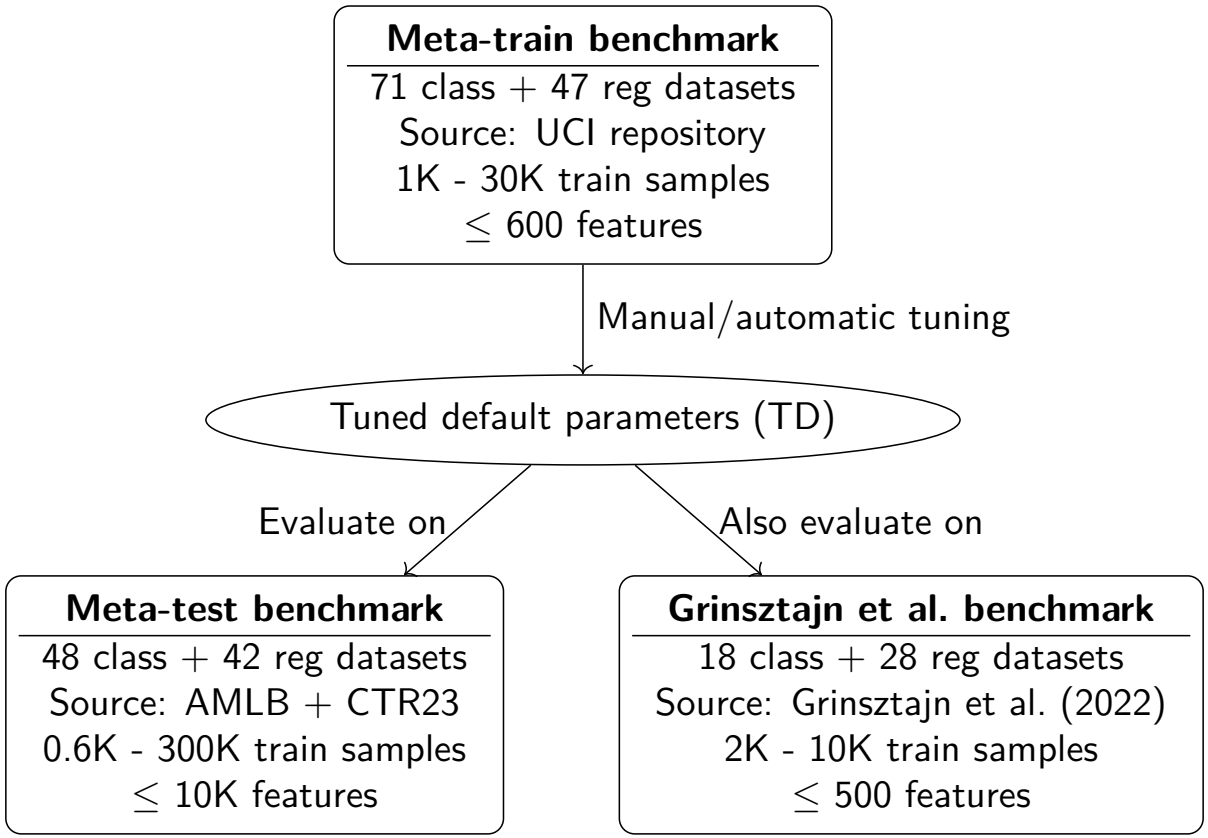
Boosted trees are fast and hard to beat

- Can we improve NNs without making them too slow?

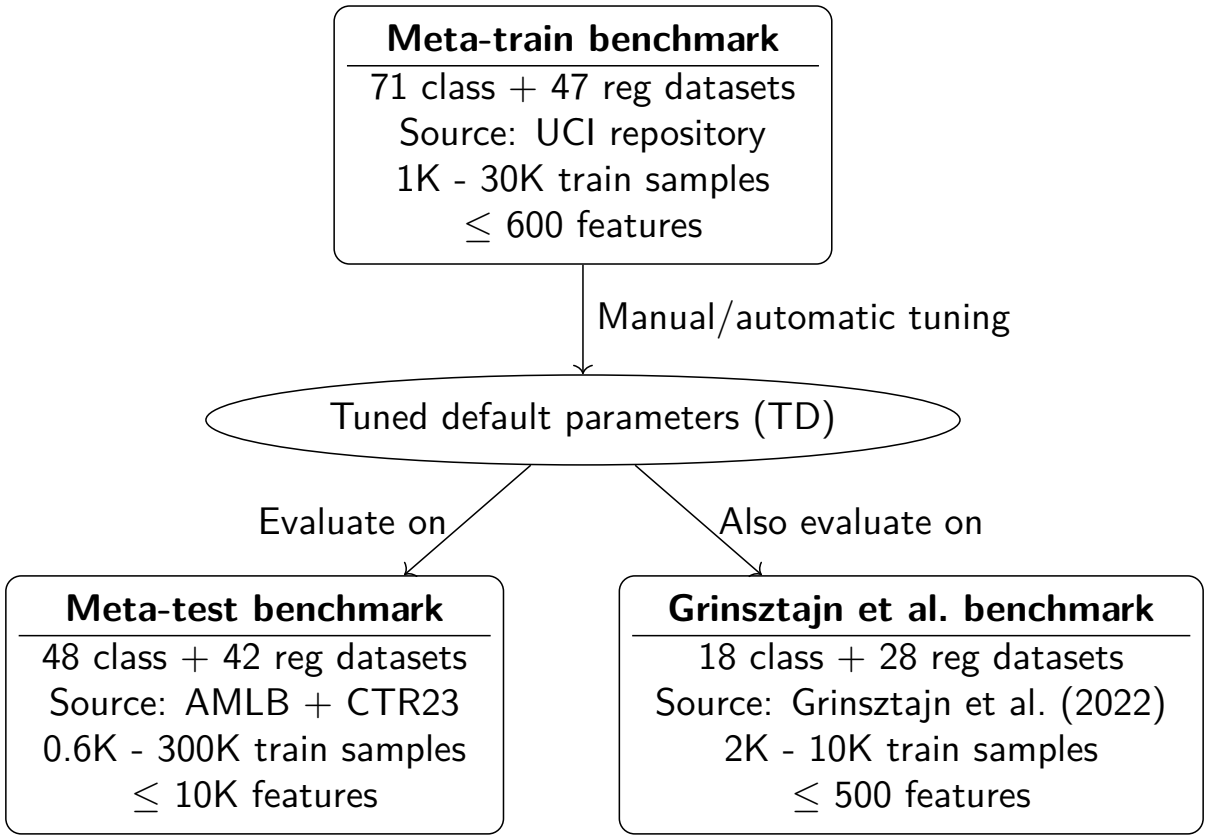
Hyperparameters are typically re-tuned on every dataset

- Can we find good fixed hyperparameters?

Meta-learning better default parameters



Meta-learning better default parameters



Target: shifted geometric mean error $= \text{geom\_mean}(\text{errors} + 0.01)$
Error metrics: $1 - \text{accuracy}$ / $1 - \text{AUROC}$ / normalized RMSE

Better defaults for boosted trees

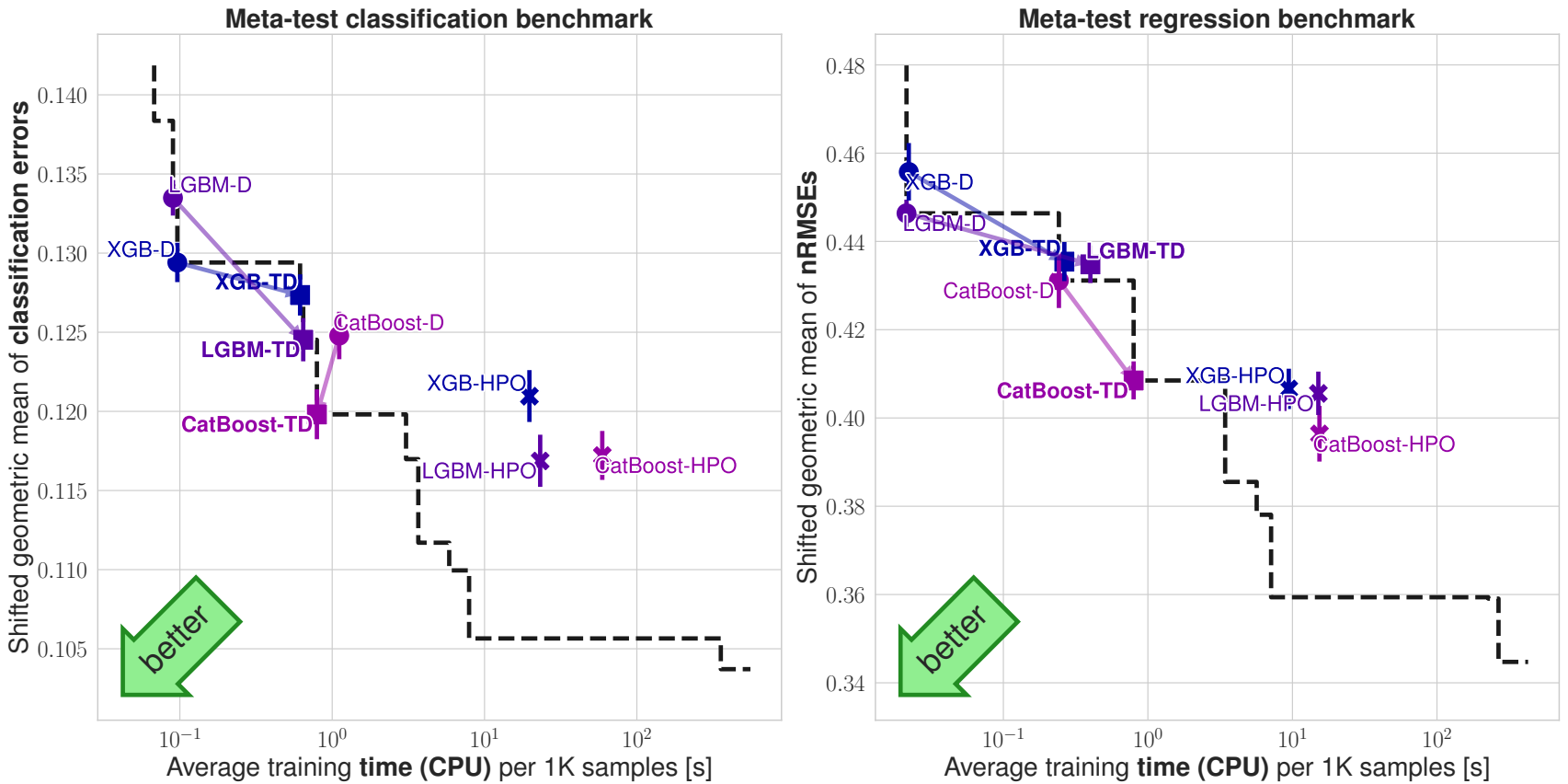


Figure: Meta-test benchmark results for the shifted geometric mean error.

RealMLP: an improved multilayer perceptron

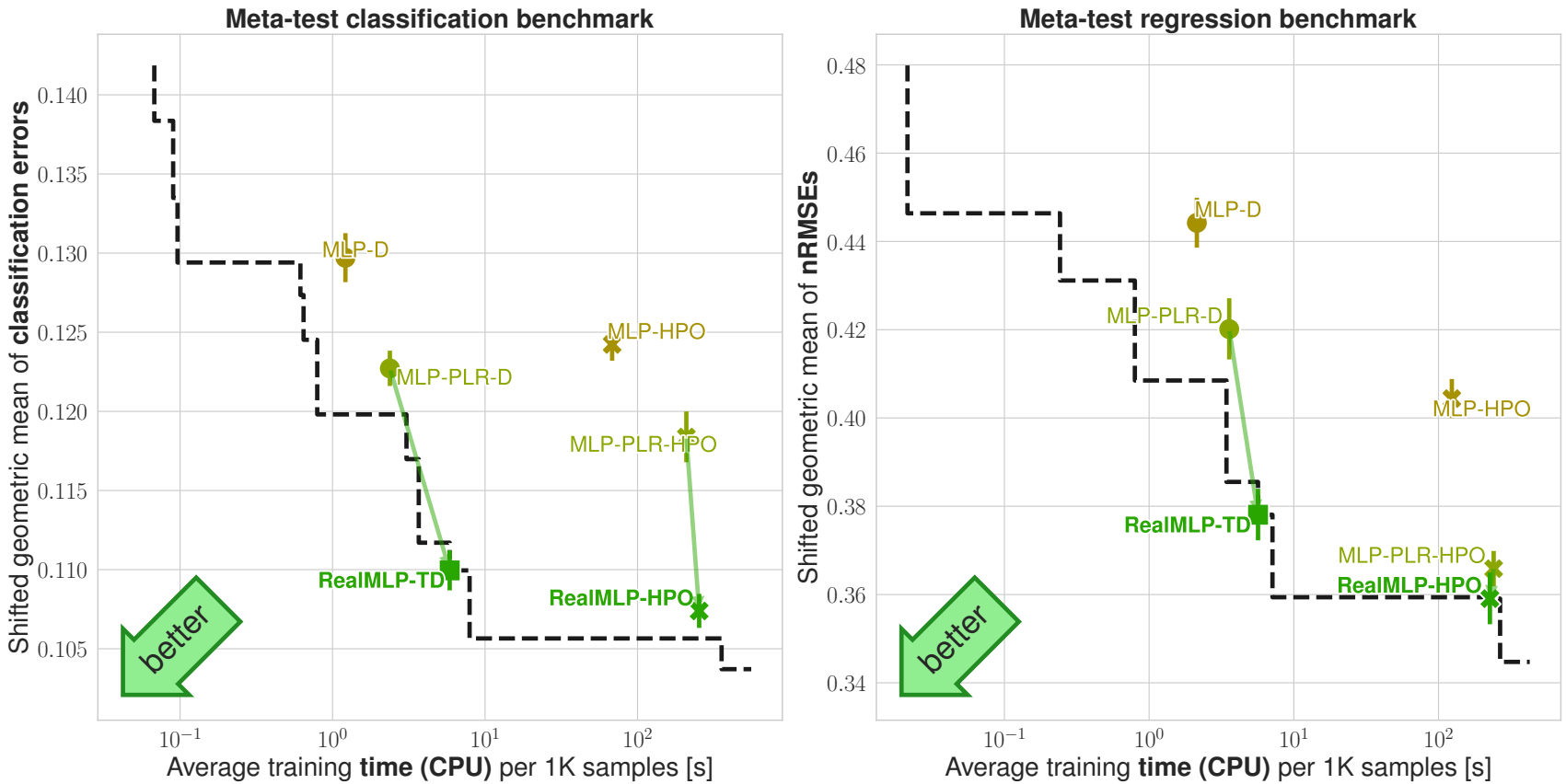


Figure: Meta-test benchmark results for the shifted geometric mean error.

Our improvements also help TabR

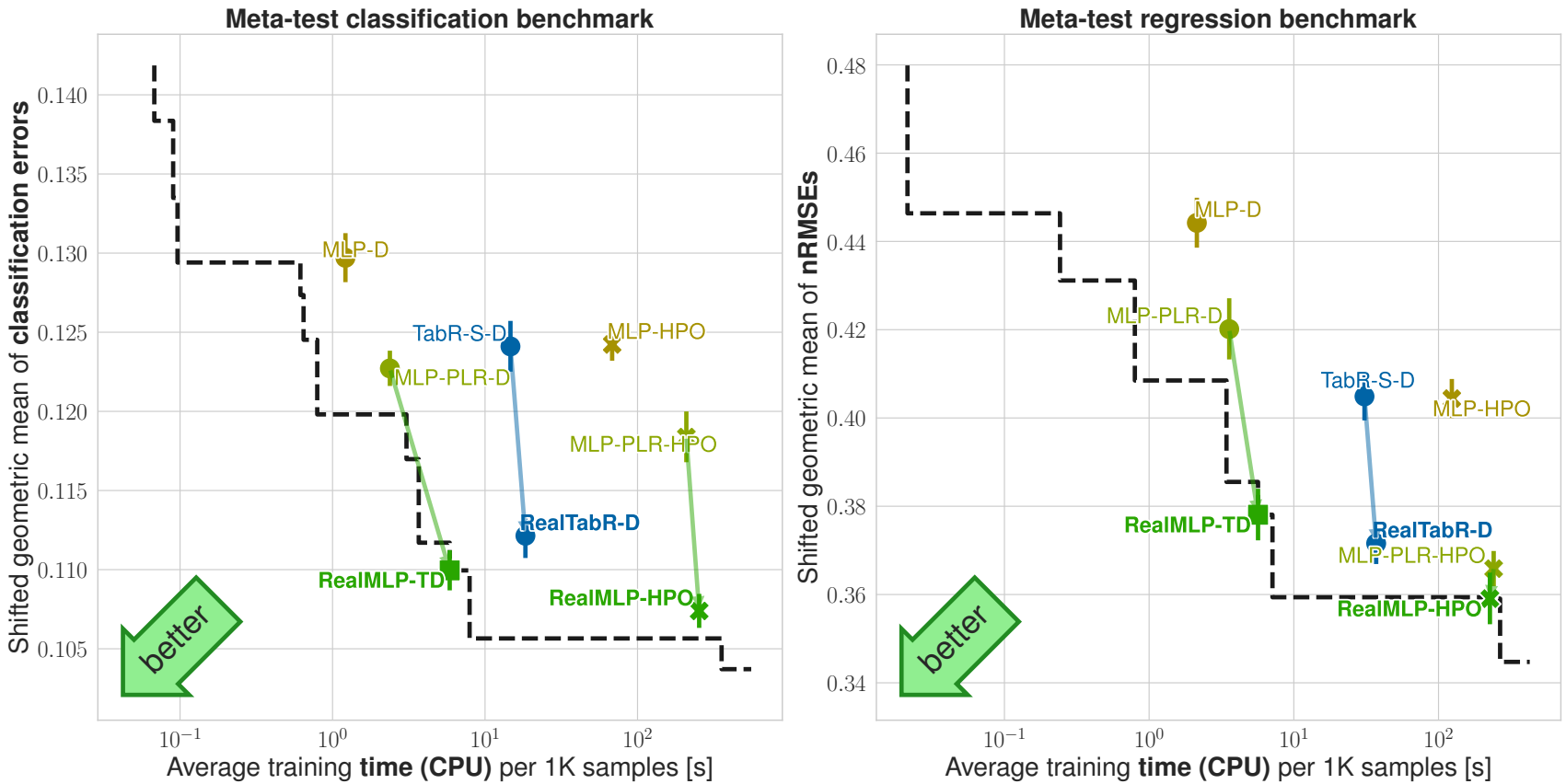


Figure: Meta-test benchmark results for the shifted geometric mean error.

RealMLP vs boosted trees: take both!

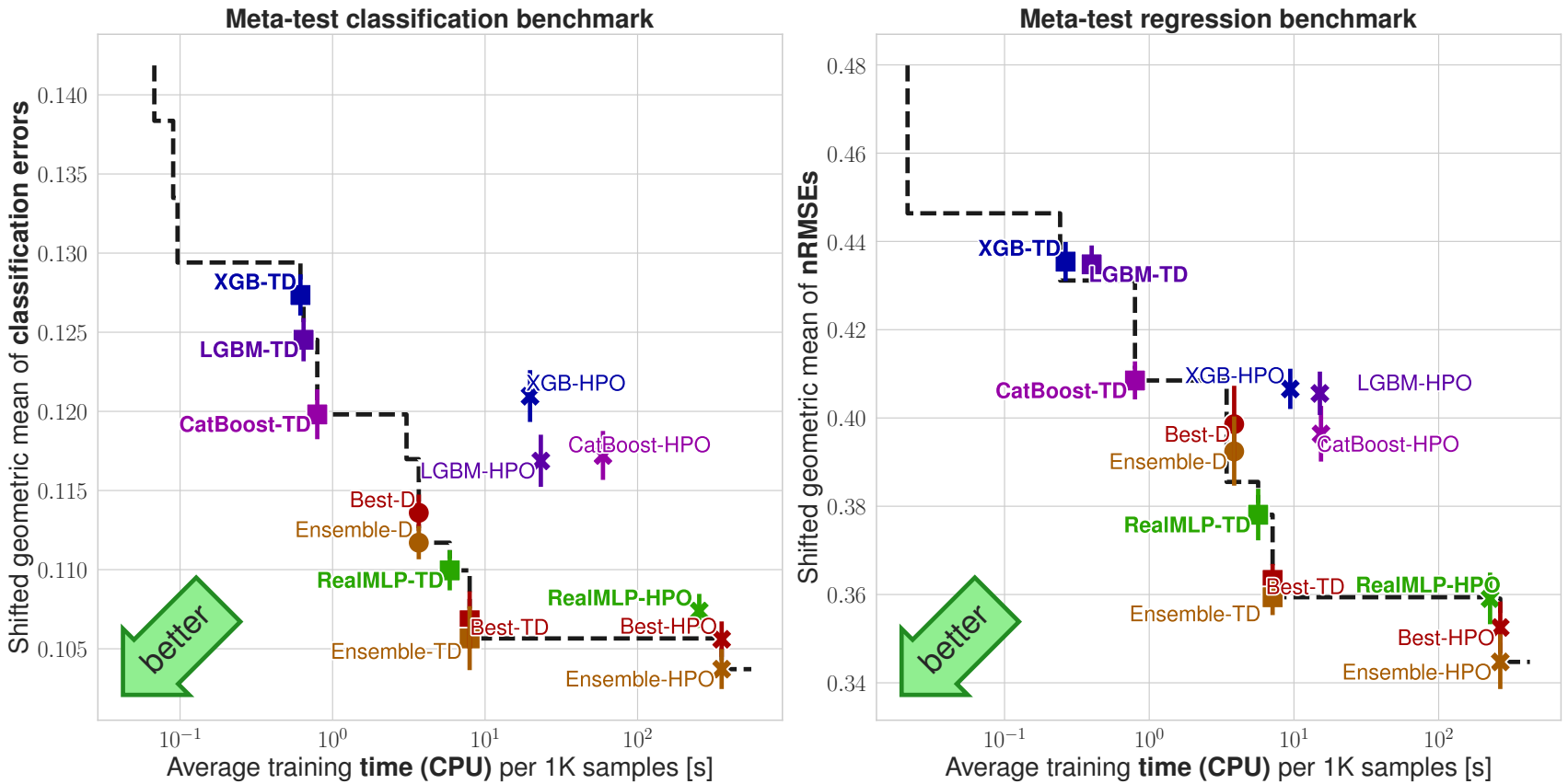
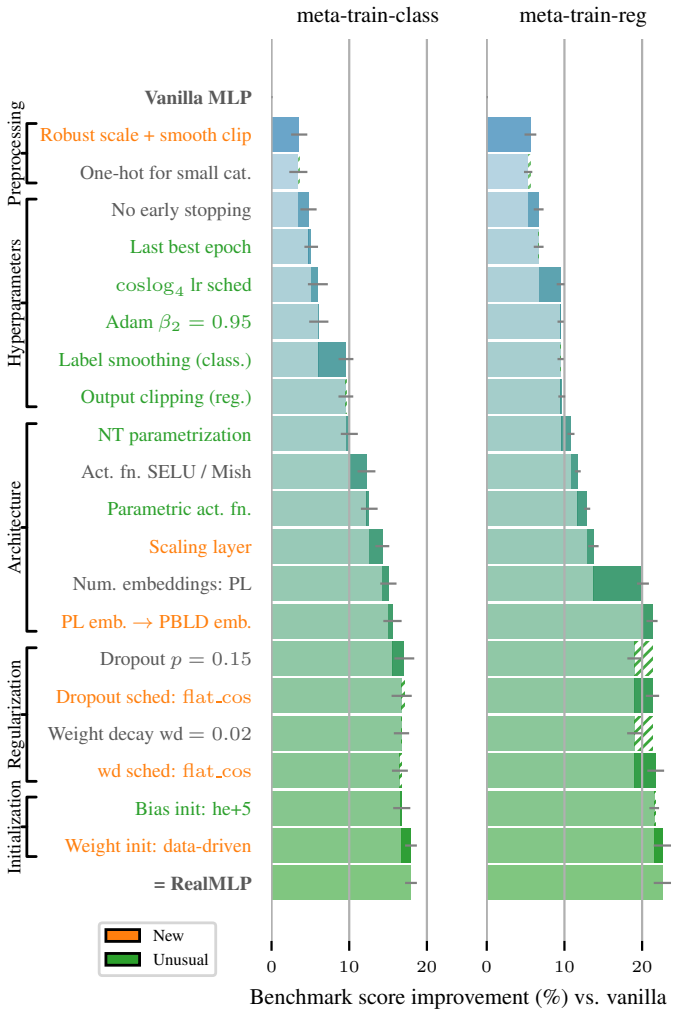
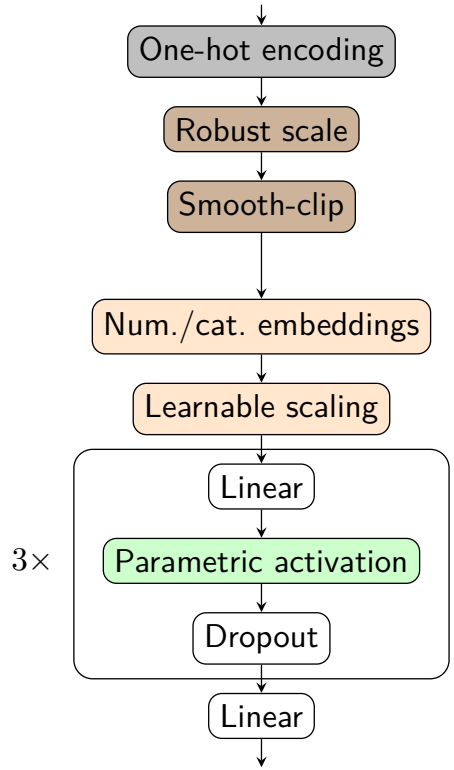
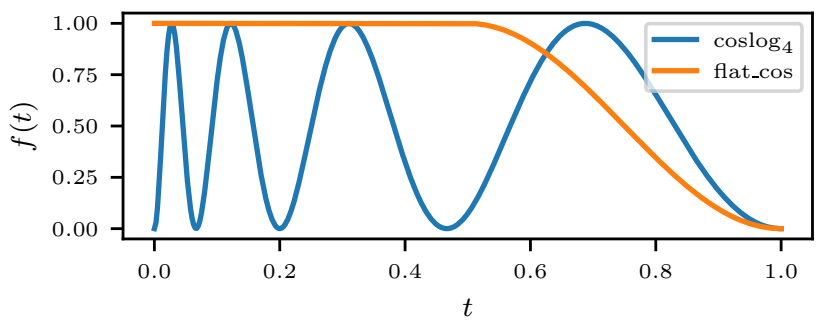


Figure: Meta-test benchmark results for the shifted geometric mean error.

Ablating NN improvements







Try it yourself!

```
pip install pytabkit
```

```python
from pytabkit import RealMLP_TD_Classifier
clf = RealMLP_TD_Classifier()  # highly configurable
clf.fit(X_train, y_train)  # works with dataframes
```

Literature I

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Neural Information Processing Systems*, 2022.