Generalization Can Emerge in Tabular Foundation Models From a Single Table

Junwei Ma[† 1], **Nour Shaheen**[2,3], **Alex Labach**[4], **Amine Mhedhbi**[2,3], **Frank Hutter**[5,6,7], **Anthony L. Caterini**[4], **Valentin Thomas**[† 4]
[1]University of Toronto
[2]Polytechnique Montréal
[3]Mila – Quebec AI Institute
[4]Layer 6 AI
[5]Prior Labs
[6]ELLIS Institute Tübingen
[7]University of Freiburg

Abstract

Deep tabular modelling increasingly relies on in-context learning where, during inference, a model receives a set of (x, y) pairs as context and predicts labels for new inputs without weight updates. We challenge the prevailing view that broad generalization here requires pre-training on large synthetic corpora (e.g., TabPFN priors) or a large collection of real data (e.g., TabDPT training datasets), discovering that a relatively small amount of data suffices for generalization. We find that simple self-supervised pre-training on just a *single* real table can produce surprisingly strong transfer across heterogeneous benchmarks. By systematically pre-training and evaluating on many diverse datasets, we analyze what aspects of the data are most important for building a Tabular Foundation Model (TFM) generalizing across domains. We then connect this to the pre-training procedure shared by most TFMs and show that the number and quality of *tasks* one can construct from a dataset is key to downstream performance.

1 Introduction

Tabular data is commonly viewed as highly diverse and heterogeneous, leading to a prevailing belief that transfer learning across different tabular domains is nearly impossible: e.g., *how could training on handwritten digits possibly transfer to estimating California housing prices?* However, recent tabular in-context learning (ICL) methods such as TABPFN [6, 7], TABDPT [9], and TABICL [12] – often termed tabular foundation models (TFMs) – have challenged this assumption.

These methods train transformers using either massive amounts of synthetic data (millions of tables generated from structured priors [6, 7, 12]) or by sampling from large collections of real tabular datasets [9], randomizing context composition and prediction targets at each step to encourage in-context generalization. All state-of-the-art tabular ICL models of which we are aware [6, 7, 9, 12, 4, 13] share a similar task construction procedure during pre-training: a dataset is either selected or created, a column of this dataset is used as the target, and a subset of the remaining columns is used as features. Inference is then performed on new datasets while keeping the pre-trained weights frozen (much like LLMs); labeled examples from the evaluation dataset are given as context and the models must predict the values of unlabeled examples. The general belief is that, much like in large language models, only a very diverse pre-training set can cover the distribution of unseen tabular tasks and enable out-of-domain generalization.

A surprising finding underpins our work: even with a *single* real-world table – such as vectorized MNIST [8], where each row represents an image flattened into pixels and each column is a pixel

[†]Correspondence: junweima2@gmail.com, vltn.thomas@gmail.com

AI for Tabular Data workshop at EurIPS 2025

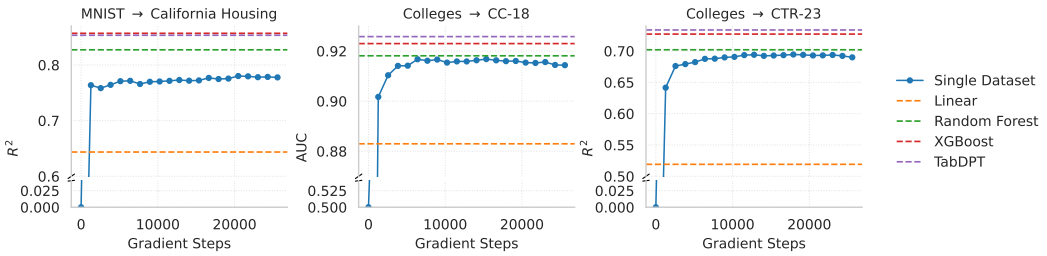


Figure 1: Transfer from a single pre-training dataset. **Left:** Training only on vectorized MNIST (treated as a table) and evaluating on CALIFORNIA HOUSING. **Middle and Right:** Training on the COLLEGES dataset and evaluating on the full CC-18 and CTR-23 evaluation suites, respectively.

value – a transformer trained with a naïve self-supervised learning (SSL) objective can still acquire generalization capabilities that transfer robustly *across domains*.

Figure 1 illustrates this: a transformer trained from scratch on only the MNIST table immediately yields strong in-context performance on structurally and semantically unrelated datasets like CALIFORNIA HOUSING [11] (predicting real estate prices from geographical and socioeconomic features). We further train on the COLLEGES dataset (7k instances, 45 features) [5], which contains demographics and institutional features, and evaluate the resulting model on the OpenML CC-18 [1] (72 datasets) and CTR-23 [3] (35 datasets) benchmarks. We find that the model learns quickly and ends up closely matching the performance of a random forest baseline. This observation stands in contrast to the prevailing assumption that only diverse, massive, or domain-matched pre-training corpora – such as those used by TABPFN and TABDPT – could enable such generalization.

In this paper, we investigate which properties of single, real-world datasets for pre-training are most linked to generalization for tabular ICL models. We find that: (1) Datasets that transfer well tend to do so universally across evaluation sets; (2) The number of features present in a dataset is a strong indicator of its generalization ability, while the number of instances matters little; and finally (3) The number of *unique tasks* a model is pre-trained on strongly affects its generalization ability.

We believe this work uncovers a very surprising phenomenon and opens the door to more investigations on how much can be learned from even limited amounts of data.

2 Experiments

2.1 Experimental Setup

We design experiments to explore the central question of what constitutes a "good dataset" such that a tabular ICL model pre-trained on it can generalize. To isolate the effects of the training datasets themselves, we fix both the model architecture and the pre-training procedure throughout most experiments. Specifically, we use the shared backbone architecture from TABPFN v1 and TABDPT, a widely adopted design for tabular ICL. As for the training objective, we use the pre-training procedure from TABDPT to train tabular ICL models from scratch using a single real dataset only.

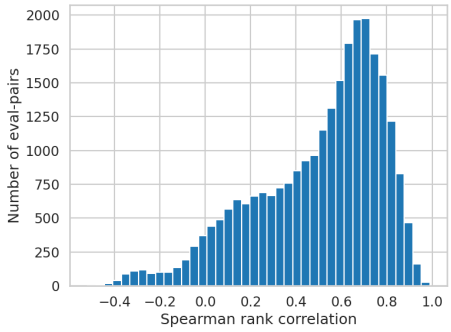
To isolate the effect of data, we choose *not* to use retrieval mechanisms (as used in TABDPT) during pre-training, although it is enabled during inference. We evaluate the generalization performance on the established benchmarks in order to easily compare with the baseline performance directly.

2.2 Are Good Pre-training Datasets Universally Good?

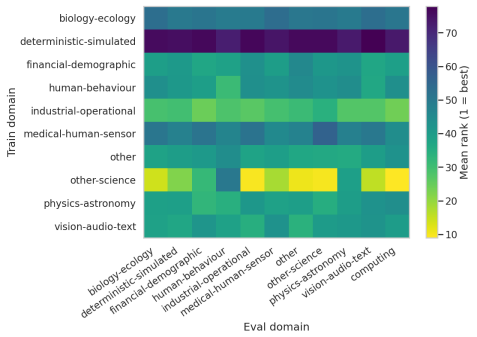
In Figure 1, we showed the surprising finding that pre-training an ICL tabular model on MNIST can lead to strong performance on other unrelated datasets. We want to investigate the following question: *"Is each pre-training dataset only good for some downstream task or does it generalize universally?"*

Thus, we pre-train models on each of $N_{\text{train}} = 88$ training datasets (a subset of the TABDPT pre-training data) separately, spanning different modalities and sizes. We evaluated each of the resulting 88 models on $N_{\text{eval}} = 107$ diverse datasets: the 72 CC-18 [1] classification datasets and the 35 CTR-23 [3] regression datasets. We took special care to ensure no datasets appear in both

the pre-training and evaluation corpora. Linear and RandomForest baselines were computed using `scikit-learn` defaults[1]. XGB is the tuned baseline from McElfresh et al. [10] and TabDPT is computed from the public repository.



(a) Ranking Correlation: The high values indicate that pre-training datasets which are good on one evaluation dataset tend to be good on others too.



(b) Domain to Domain Transfer: We do not observe stronger transfer when the pre-training and evaluation domain are shared.

Figure 2: Universality of the dataset quality. **Figure 2a:** We compute the rank of training sets for each evaluation dataset and then plot a histogram of the Spearman correlation between the ranks on all pairs of evaluation datasets. Most evaluation datasets have very correlated ranks. **Figure 2b:** We group the training and evaluation datasets into distinct domains and plot the density map for average ranks (lower is better). Training and evaluation pairs from the same domain do not appear to transfer better than ones from different domains.

In Figure 2a, for each evaluation dataset we rank the pre-training datasets based on the performance obtained on the evaluation task, leading to N_{eval} rankings. We then gather all pairs of rankings and compute the Spearman correlation. Higher correlations indicate that the relative ranking of pre-training datasets is stable. Since the histogram has much more mass to the right, it indicates that if a dataset is good for one evaluation dataset it tends to be good for others too, *even across vastly different evaluation tasks*.

Next, we investigate whether the *domain* of the single training dataset tends to generalize better to evaluation datasets within the same domain. For this purpose, we categorize the pre-training and evaluation datasets by manual inspection. Our protocol is the following: we group the evaluation datasets by domain and take the final performance (either accuracy for classification or correlation for regression) for each single training dataset. We then rank the training datasets from best to worst for each domain. Finally we aggregate over the domains of the training datasets. If datasets from, e.g., Domain A are expected to transfer, on average, better to Domain A, we would expect a strong diagonal dominance in this matrix. However, in Figure 2b we can see this is absolutely not the case. Domains do not seem to transfer better to the same domain. Rather, certain domains appear to be better than others across the board (e.g., `other-science` is great, `deterministic-simulated` is not).

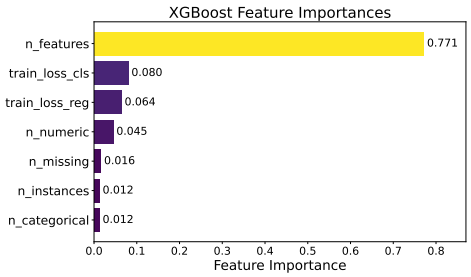
2.3 What Constitutes a Good Dataset?

To understand which properties of a training dataset yield successful generalization, we conduct a meta-analysis using the previous pre-training corpus of N_{train} datasets. For each pre-training dataset, we collect a set of descriptive meta-features, including: the number of features, number of instances, number of categorical features, number of numeric features, amount of missing values, and final pre-training losses for both classification and regression heads. These meta-features are split into meta-train (80%) and meta-test (20%) sets.

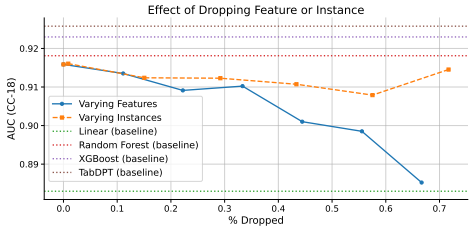
We then train an XGBoost [2] regressor to predict a dataset's average downstream generalization score (chosen as the average of correlation for regression tasks and accuracy for classification tasks) using these meta-features. The resulting model achieves an R^2 of 0.67, indicating that roughly two-thirds of the variance in generalization can be reproduced from straightforward dataset properties alone.

[1] except with 1000 maximum iterations for logistic regression instead of 100

Notably, when evaluating feature importance (see Figure 3a), the number of features emerges as by far the strongest predictor of downstream transfer performance. In contrast, the number of instances – often presumed critical in classical settings – shows negligible contribution to the XGBoost's predictive power. This suggests that the richness of the feature space is far more important than purely the dataset size for tabular ICL-based generalization. We validate this on Figure 3b by removing a fraction of either rows or columns on the COLLEGES dataset, showing that features tend to matter much more than instances. After dropping about 70% of features, the model performance drops to a similar level as a linear model, while dropping more than 70% of instances still allows the model to maintain performance close to a random forest. We see a similar phenomenon across many other datasets.



(a) Feature importance from an XGBoost model trained to predict the average performance across 107 downstream tasks from dataset metadata.



(b) Column vs. Row Importance: We analyze the downstream performance when removing either features or instances from the `College` dataset.

Figure 3: Not all cells in a table are made equal: the number of features matters much more than the number of instances as pre-training dataset for tabular ICL models.

2.4 Training on Many Good Tasks Unlocks Generalization

In Figure 3b, we observed that dropping features caused a strong drop in performance. To understand why, we tie this observation back to the randomized procedure used for pre-training. As mentioned, a column is selected as target while a subset of other columns are used as features. We call this a *task*. For instance, a dataset with a fixed supervised target is one task, while using a randomized target and features can have $\mathcal{O}(k\ 2^k)$ tasks for k columns. Thus the role of columns may be more important than the role of rows as it is directly tied to the number of tasks or relationships between features the model is exposed to during training.

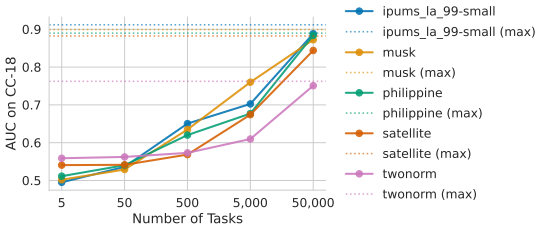


Figure 4: Downstream AUC as a function of the number of unique tasks used during training. More tasks during pre-training consistently leads to better transfer. This demonstrates that both the amount of tasks and the quality of tasks drive the generalization performance.

We disentangle the impact of the number of tasks from the number of features in Figure 4. We do not change the amount of instances or features in a given dataset but instead we vary the number of tasks the model is allowed to pre-train on. To be precise, we select N tasks using allowed permutations (one target, the remaining columns being either masked or used as features) and vary N for several datasets.

Our results in Figure 4 show a striking pattern: across all datasets, increasing the number of tasks from 5 to tens of thousands drives the performance from barely above random (AUC $= 0.5$) to good performance. This underscores that while there are important inherent differences across datasets and that some datasets are significantly better than others for pre-training, the main factor explaining generalization across domains is the number of tasks a model was trained on during pre-training.

In other words, it is not so much the amount of data (in tokens or bytes of memory) of the pre-training set that is key, but the amount of tasks / relationships between features and targets trained on that enable a tabular ICL model to generalize. While we believe task diversity and quality are other key elements, we leave a deeper investigation into these factors for future work.

References

[1] Bernd Bischl, Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael Gomes Mantovani, Jan van Rijn, and Joaquin Vanschoren. OpenML benchmarking suites. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

[2] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SigKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[3] Sebastian Felix Fischer, Matthias Feurer, and Bernd Bischl. OpenML-CTR23 – A curated tabular regression benchmarking suite. In *AutoML Conference (Workshop)*, 2023.

[4] Josh Gardner, Juan C Perdomo, and Ludwig Schmidt. Large scale transfer learning for tabular data via language modeling. In *Advances in Neural Information Processing Systems*, 2024.

[5] Pieter Gijsbers. Colleges. OpenML Dataset 42727, 2020. URL `https://www.openml.org/d/42727`. Accessed: 2025-10-21.

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

[7] 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.

[8] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[9] Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Hamidreza Kamkari, Alex Labach, Jesse C. Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L. Caterini, and Maksims Volkovs. TabDPT: Scaling tabular foundation models on real data. In *Advances in Neural Information Processing Systems*, 2025.

[10] Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? In *Advances in Neural Information Processing Systems*, 2023.

[11] R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997.

[12] Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *International Conference on Machine Learning*, 2025.

[13] Marco Spinaci, Marek Polewczyk, Maximilian Schambach, and Sam Thelin. ConTextTab: A semantics-aware tabular in-context learner. In *1st ICML Workshop on Foundation Models for Structured Data*, 2025.

A Appendix / supplemental material

A.1 Impact of Pre-training Dataset on CC-18 and CTR-23 Performance

Tables1, 2 report the performance of TabDPT when pre-trained on different datasets and evaluated on two benchmark suites: **CC-18** for classification (measured by AUC) and **CTR-23** for regression (measured by R^2). Each table lists the results for all pre-training datasets, along with their number of instances, their number of features, and their domain. The results are sorted in descending order of performance, highlighting which pre-training datasets achieve better performance on each benchmark.

Table 1: AUC scores on CC-18 for TabDPT pre-trained on various datasets, sorted in descending order of AUC.

Dataset Name	AUC on CC-18	# Instances	# Features	Dataset Domain
colleges	0.916	7063	45	other
ipums_la_97-small	0.914	7019	61	financial-demographic
cardiotocography	0.913	2126	36	medical-human-sensor
ipums_la_98-small	0.912	7485	56	financial-demographic
ipums_la_99-small	0.912	8844	57	financial-demographic
volkert	0.910	58310	181	other
KDDCup09_appetency	0.910	50000	231	human-behaviour
APSFailure	0.909	76000	171	industrial-operational
KDDCup09_churn	0.909	50000	231	industrial-operational
Census-Income	0.907	299285	42	financial-demographic
gas-drift	0.906	13910	129	other-science
gas-drift-different-concentrations	0.903	13910	130	other-science
kick	0.903	72983	33	industrial-operational
MiniBooNE	0.902	130064	51	physics-astronomy
christine	0.901	5418	1637	other
road-safety	0.901	111762	33	human-behaviour
dilbert	0.900	10000	2001	other
musk	0.899	6598	168	other-science
covertype	0.898	581012	55	biology-ecology
PizzaCutter3	0.894	1043	38	other
scene	0.894	2407	300	vision-audio-text
dionis	0.893	416188	61	other
eye_movements	0.893	10936	28	medical-human-sensor
helena	0.892	65196	28	other
philippine	0.890	5832	309	other
heloc	0.889	10000	23	financial-demographic
SpeedDating	0.888	8378	121	human-behaviour
one-hundred-plants-texture	0.887	1599	65	biology-ecology
kdd_internet_usage	0.886	10108	69	financial-demographic
fbis.wc	0.886	2463	2001	vision-audio-text
jannis	0.883	83733	55	other
colleges_usnews	0.883	1302	34	other
Satellite	0.882	5100	37	physics-astronomy
porto-seguro	0.881	595212	58	human-behaviour
PieChart3	0.881	1077	38	other
sylva_agnostic	0.879	14395	217	biology-ecology
jasmine	0.874	2984	145	other
sylvine	0.874	5124	21	other
elevators	0.871	16599	19	other
guillermo	0.870	20000	4297	other
JapaneseVowels	0.868	9961	15	vision-audio-text
ada	0.867	4147	49	other
cjs	0.865	2796	35	biology-ecology
higgs	0.865	98050	29	physics-astronomy

Table 1: AUC scores on CC-18 for TabDPT pre-trained on various datasets, sorted in descending order of AUC.

Dataset Name	AUC on CC-18	# Instances	# Features	Dataset Domain
ada_agnostic	0.862	4562	49	financial-demographic
shuttle	0.854	58000	10	physics-astronomy
house_16H	0.850	22784	17	financial-demographic
pol	0.849	10082	27	industrial-operational
okcupid-stem	0.840	50789	20	human-behaviour
eeg-eye-state	0.839	14980	15	medical-human-sensor
analcatdata_halloffame	0.833	1340	17	other
page-blocks	0.828	5473	11	vision-audio-text
default-of-credit-card-clients	0.825	13272	21	financial-demographic
spoken-arabic-digit	0.825	263256	15	vision-audio-text
albert	0.823	425240	79	other
MagicTelescope	0.820	19020	12	physics-astronomy
mushroom	0.820	8124	23	biology-ecology
magic	0.819	19020	11	physics-astronomy
Click_prediction_small	0.816	39948	12	human-behaviour
pbcseq	0.815	1945	19	medical-human-sensor
fabert	0.814	8237	801	other
sf-police-incidents	0.807	2215023	9	human-behaviour
credit	0.807	16714	11	financial-demographic
ldpa	0.802	164860	8	medical-human-sensor
airlines	0.794	539383	8	industrial-operational
artificial-characters	0.786	10218	8	deterministic-simulated
ada_prior	0.768	4562	15	financial-demographic
madeline	0.765	3140	260	other
twonorm	0.762	7400	21	deterministic-simulated
house_8L	0.758	22784	9	financial-demographic
Diabetes130US	0.756	71090	8	medical-human-sensor
yeast	0.748	1269	9	biology-ecology
pollen	0.748	3848	6	biology-ecology
hill-valley	0.744	1212	101	deterministic-simulated
walking-activity	0.737	149332	5	medical-human-sensor
compas-two-years	0.737	4966	12	human-behaviour
visualizing_soil	0.727	8641	5	biology-ecology
poker-hand	0.706	1025009	11	deterministic-simulated
puma8NH	0.698	8192	9	deterministic-simulated
kropt	0.688	28056	7	deterministic-simulated
chess	0.686	28056	7	deterministic-simulated
kr-vs-k	0.675	28056	7	deterministic-simulated
analcatdata_supreme	0.668	4052	8	other

Table 2: R^2 scores on CTR-23 for TabDPT pre-trained on various datasets, sorted in descending order of R^2.

Dataset Name	R^2 on CTR-23	# Instances	# Features	Dataset Domain
colleges	0.694	7063	45	other
covertype	0.679	581012	55	biology-ecology
ipums_la_99small	0.677	8844	57	financial-demographic
ipums_la_98small	0.672	7485	56	financial-demographic
kick	0.668	72983	33	industrial-operational
Census-Income	0.666	299285	42	financial-demographic
gas-drift-different-concentrations	0.666	13910	130	other-science
volkert	0.663	58310	181	other

Table 2: R^2 scores on CTR-23 for TabDPT pre-trained on various datasets, sorted in descending order of R^2.

Dataset Name	R^2 on CTR-23	# Instances	# Features	Dataset Domain
ipums_la_97small	0.662	7019	61	financial-demographic
KDDCup09_appetency	0.661	50000	231	human-behaviour
MiniBooNE	0.661	130064	51	physics-astronomy
APSFailure	0.660	76000	171	industrial-operational
KDDCup09_churn	0.659	50000	231	industrial-operational
gas-drift	0.656	13910	129	other-science
cardiotocography	0.656	2126	36	medical-human-sensor
scene	0.652	2407	300	vision-audio-text
jannis	0.652	83733	55	other
dilbert	0.652	10000	2001	other
philippine	0.652	5832	309	other
helena	0.650	65196	28	other
one-hundred-plants-texture	0.644	1599	65	biology-ecology
PizzaCutter3	0.644	1043	38	other
eye_movements	0.642	10936	28	medical-human-sensor
sylva_agnostic	0.642	14395	217	biology-ecology
dionis	0.641	416188	61	other
colleges_usnews	0.640	1302	34	other
christine	0.638	5418	1637	other
road-safety	0.636	111762	33	human-behaviour
heloc	0.633	10000	23	financial-demographic
PieChart3	0.633	1077	38	other
JapaneseVowels	0.626	9961	15	vision-audio-text
higgs	0.625	98050	29	physics-astronomy
SpeedDating	0.624	8378	121	human-behaviour
sylvine	0.620	5124	21	other
elevators	0.619	16599	19	other
shuttle	0.613	58000	10	physics-astronomy
musk	0.612	6598	168	other-science
jasmine	0.608	2984	145	other
analcatdata_halloffame	0.606	1340	17	other
Satellite	0.605	5100	37	physics-astronomy
page-blocks	0.604	5473	11	vision-audio-text
pol	0.602	10082	27	industrial-operational
eeg-eye-state	0.595	14980	15	medical-human-sensor
house_16H	0.592	22784	17	financial-demographic
guillermo	0.581	20000	4297	other
porto-seguro	0.576	595212	58	human-behaviour
spoken-arabic-digit	0.562	263256	15	vision-audio-text
cjs	0.558	2796	35	biology-ecology
fbis.wc	0.553	2463	2001	vision-audio-text
pbcseq	0.551	1945	19	medical-human-sensor
ada	0.546	4147	49	other
albert	0.543	425240	79	other
ada_agnostic	0.542	4562	49	financial-demographic
artificial-characters	0.541	10218	8	deterministic-simulated
MagicTelescope	0.540	19020	12	physics-astronomy
kdd_internet_usage	0.536	10108	69	financial-demographic
mushroom	0.533	8124	23	biology-ecology
magic	0.528	19020	11	physics-astronomy
ldpa	0.510	164860	8	medical-human-sensor
okcupid-stem	0.510	50789	20	human-behaviour
default-of-credit-card-clients	0.496	13272	21	financial-demographic
ada_prior	0.490	4562	15	financial-demographic

Table 2: R^2 scores on CTR-23 for TabDPT pre-trained on various datasets, sorted in descending order of R^2.

Dataset Name	**R^2 on CTR-23**	**# Instances**	**# Features**	**Dataset Domain**
credit	0.476	16714	11	financial-demographic
madeline	0.465	3140	260	other
visualizing_soil	0.459	8641	5	biology-ecology
compas-two-years	0.452	4966	12	human-behaviour
fabert	0.446	8237	801	other
Click_prediction_small	0.430	39948	12	human-behaviour
airlines	0.421	539383	8	industrial-operational
sf-police-incidents	0.410	2215023	9	human-behaviour
yeast	0.405	1269	9	biology-ecology
house_8L	0.384	22784	9	financial-demographic
hill-valley	0.383	1212	101	deterministic-simulated
pollen	0.329	3848	6	biology-ecology
twonorm	0.322	7400	21	deterministic-simulated
Diabetes130US	0.313	71090	8	medical-human-sensor
walking-activity	0.312	149332	5	medical-human-sensor
analcatdata_supreme	0.296	4052	8	other
kropt	0.290	28056	7	deterministic-simulated
kr-vs-k	0.276	28056	7	deterministic-simulated
chess	0.275	28056	7	deterministic-simulated
puma8NH	0.271	8192	9	deterministic-simulated
poker-hand	0.263	1025009	11	deterministic-simulated