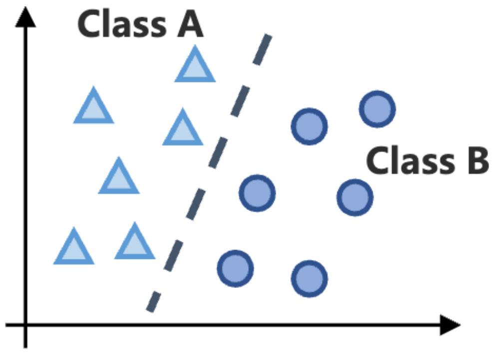
TabICL: A Tabular Foundation Model for In-Context Learning on Large Data

Jingang QU, David Holzmüller, Gaël Varoquaux, Marine Le Morvan

ICML 2025

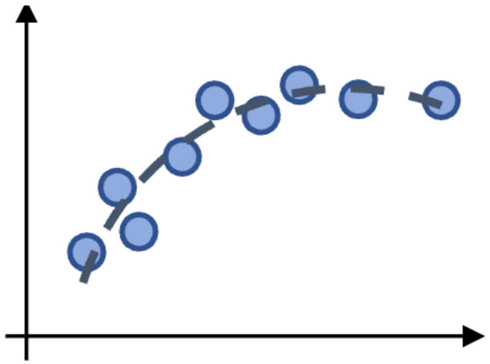
Currency	Amount	Card type	Age	*Fraud*
USD	3497.74	Debit	55	Yes
EUR	1121.53	Prepaid	45	No
CNY	2867.57	Credit	31	No
USD	4100.37	Debit	59	?

Classification



- Heterogenous features (numerical, categorical, ...)
- Missing values, uninformative features, and outliers
- Lack of spatial or sequential relationships

Regression



In-context Learning with Transformer

[Tom et al. 2020]

In-context examples

$7 + 2 = \ \ 9$	gaot => goat
$1 + 0 = \ \ 1$	sanke => snake
$3 + 4 = \ \ 7$	brid => bird
$5 + 9 = 14$	fsih => fish
$8 + 4 = 12$	dcuk => duck

Query

$9 + 8 = \ ?$	cmihp => ?

deepseek

Gemini

The pattern here involves rearranging the jumbled letters to form the correct name of an animal. So, cmihp ⇒ chimp 🐒

$X \in \mathrm{R}^{n \times m}$ $\quad$ $y \in \mathrm{R}^{n \times 1}$

$\mathcal{D}_{\text{train}}$

Columns (Features)

Rows (Samples)

x_{11}	x_{12}	$\cdots$	x_{1m}	=>	y_1
x_{21}	x_{22}	$\cdots$	x_{2m}		y_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$		$\cdots$
x_{n1}	x_{n2}	$\cdots$	x_{nm}		y_n

Target

Test

x_1^*	x_2^*	$\cdots$	x_m^*	=>	?

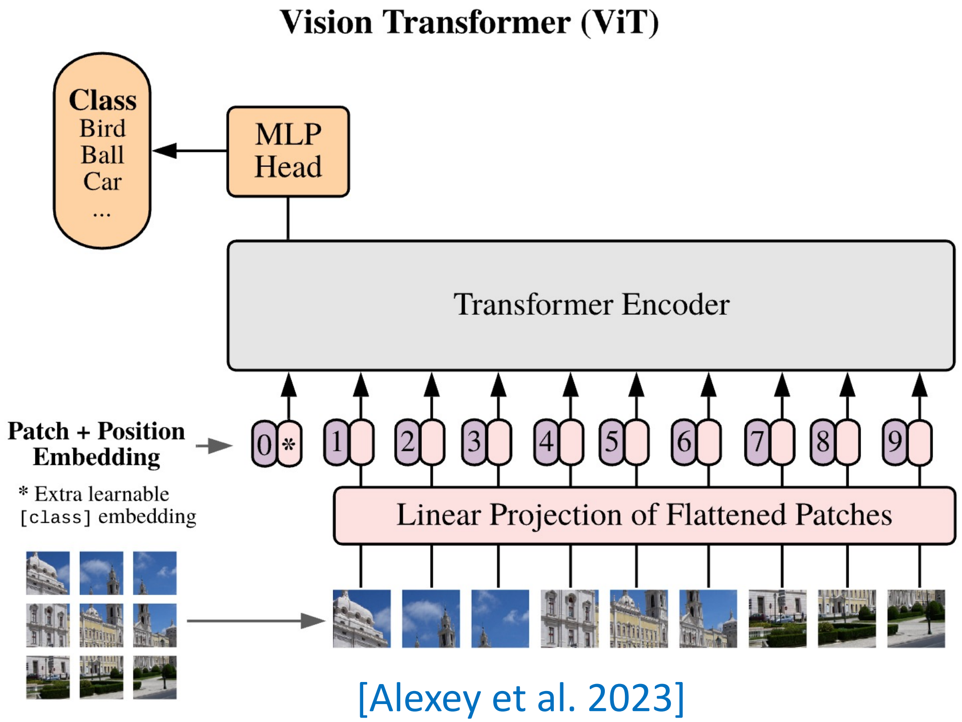
Transformer

$$p(y^* \mid x^*, \mathcal{D}_{\text{train}}; \theta)$$

Tokenization

Large Language Model

Today, I will present TabICL at the kick-off of Soda.



[Alexey et al. 2023]

$X \in \mathrm{R}^{n\times m}$ $\quad$ $y \in \mathrm{R}^{n\times 1}$

Rows (Samples)	Columns (Features)				Target
	x_{11}	x_{12}	$\dots$	x_{1m}	y_1
	x_{21}	x_{22}	$\dots$	x_{2m}	y_2
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	x_{n1}	x_{n2}	$\dots$	x_{nm}	y_n

$\{ (x_1, y_1), \dots, (x_n, y_n) \}, \; x^*$

Tabular Prior-Data Fitted Network (TabPFN)

[Hollmann et al. 2023]

$X \in \mathrm{R}^{n \times m}$

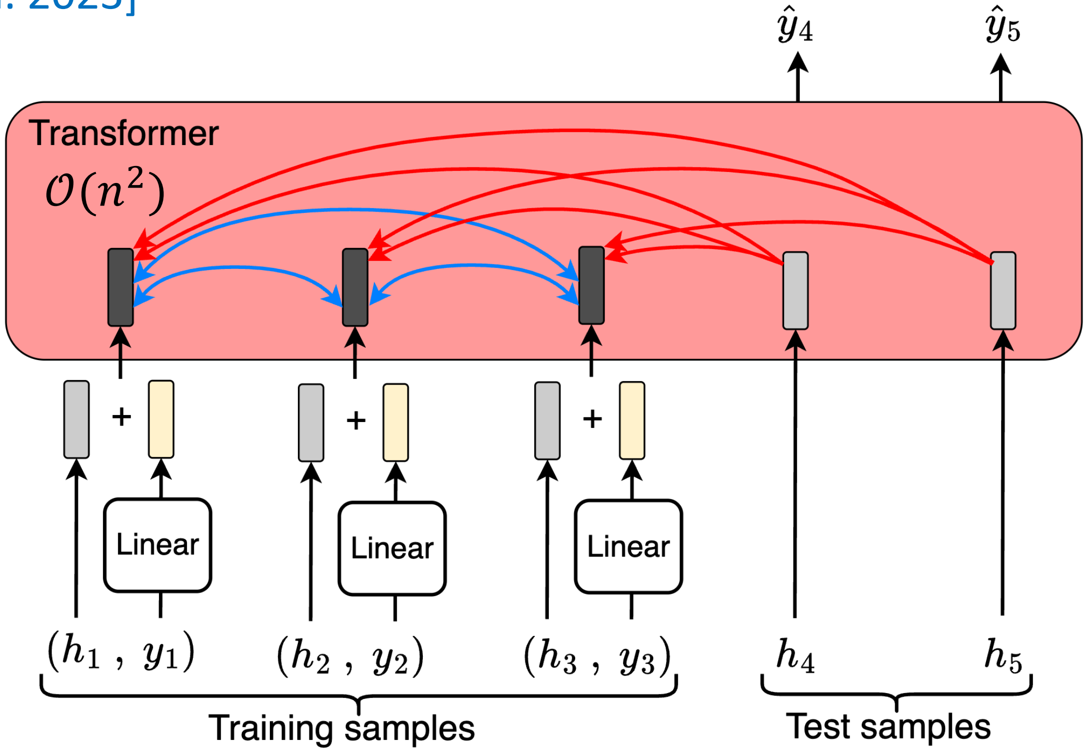
Columns (Features)

Rows (Samples)

$$\begin{array}{|c|c|c|c|} \hline x_{11} & x_{12} & \cdots & x_{1m} \\ \hline x_{21} & x_{22} & \cdots & x_{2m} \\ \hline \cdots & \cdots & \cdots & \cdots \\ \hline x_{n1} & x_{n2} & \cdots & x_{nm} \\ \hline \end{array}$$

- Each row is a token
- $h_i = \text{Linear}(\text{row}_i)$
- Attention between h_i with $\mathcal{O}(n^2)$

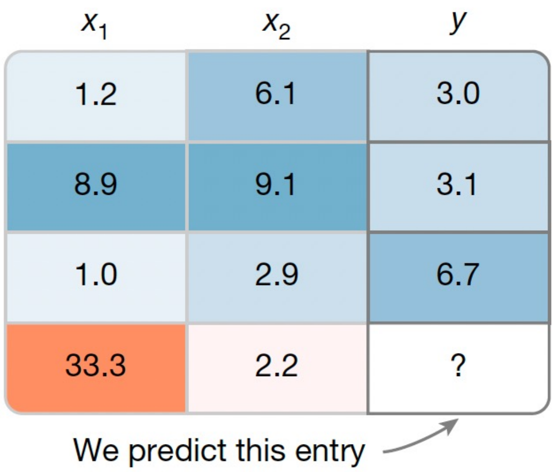
Complexity => $\mathcal{O}(n^2)$



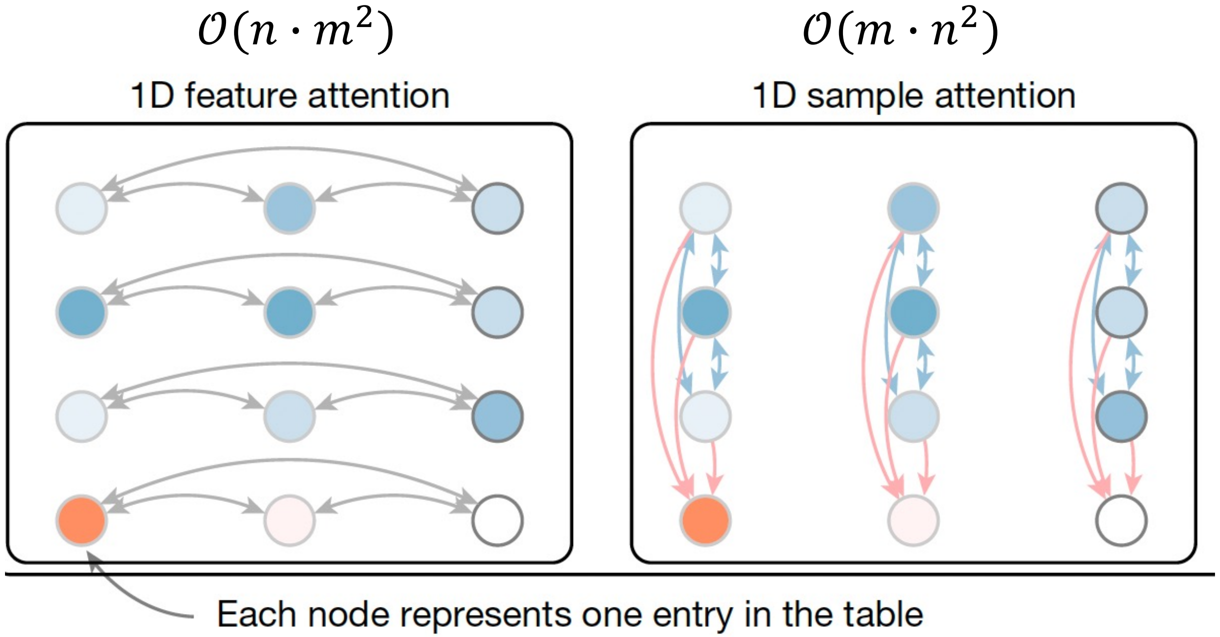
	Pretraining			*Inference*		
	# of samples	# of features	# of classes	# of samples	# of features	# of classes
TabPFN	1,024	100	10	3,000	100	10

TabPFNv2

[Hollmann et al. 2025]



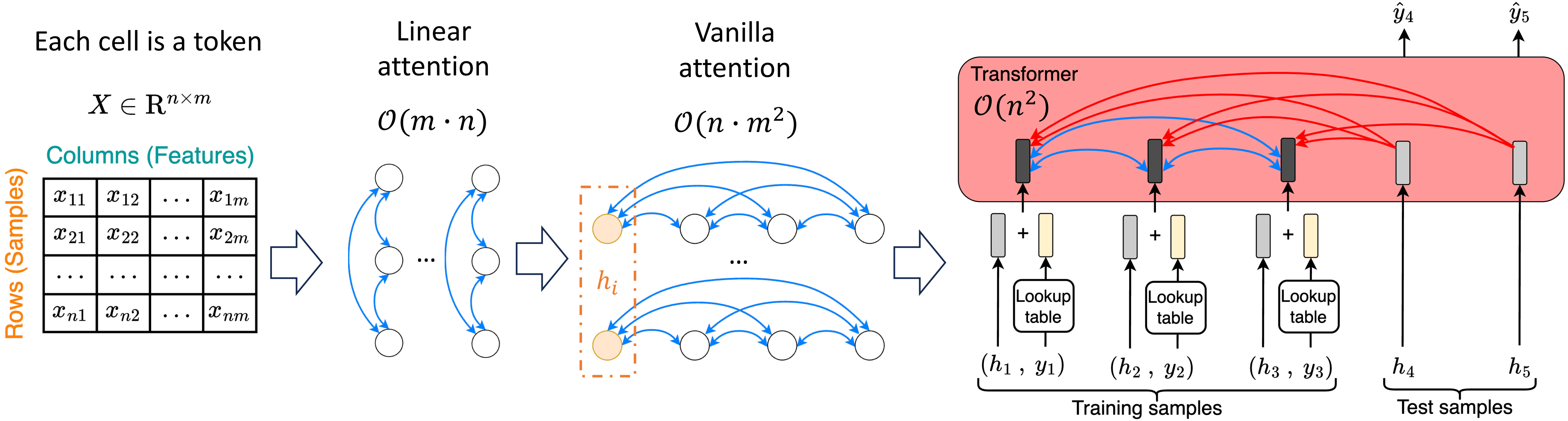
Each cell is a token



Complexity => $\mathcal{O}(n \cdot m^2) + \mathcal{O}(m \cdot n^2)$

	Pretraining			***Inference***		
	# of samples	# of features	# of classes	# of samples	# of features	# of classes
TabPFN	1,024	100	10	3,000	100	10
TabPFNv2	2,048	160	10	10,000	500	10

TabICL: A Tabular Foundation Model for Large Data



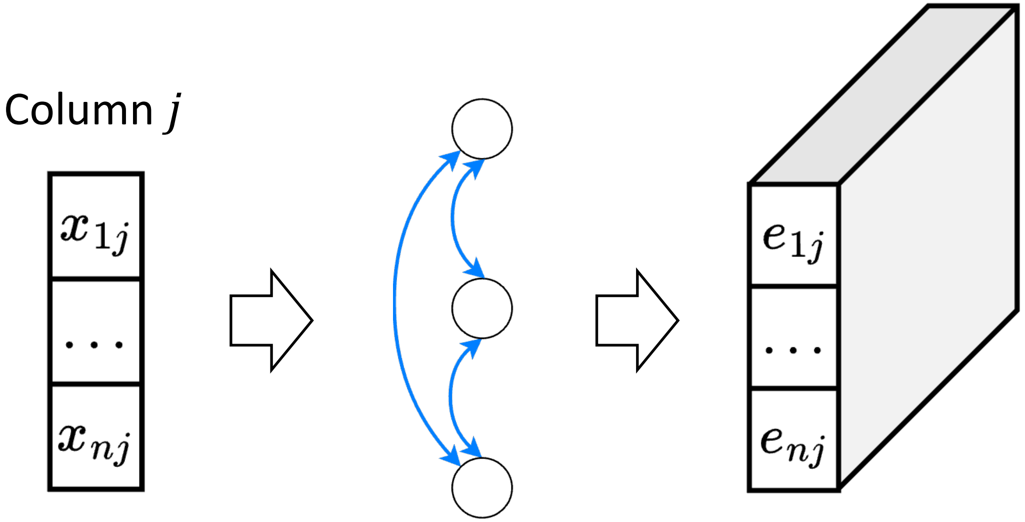
Complexity => $\mathcal{O}(n \cdot m^2) + \mathcal{O}(n^2)$

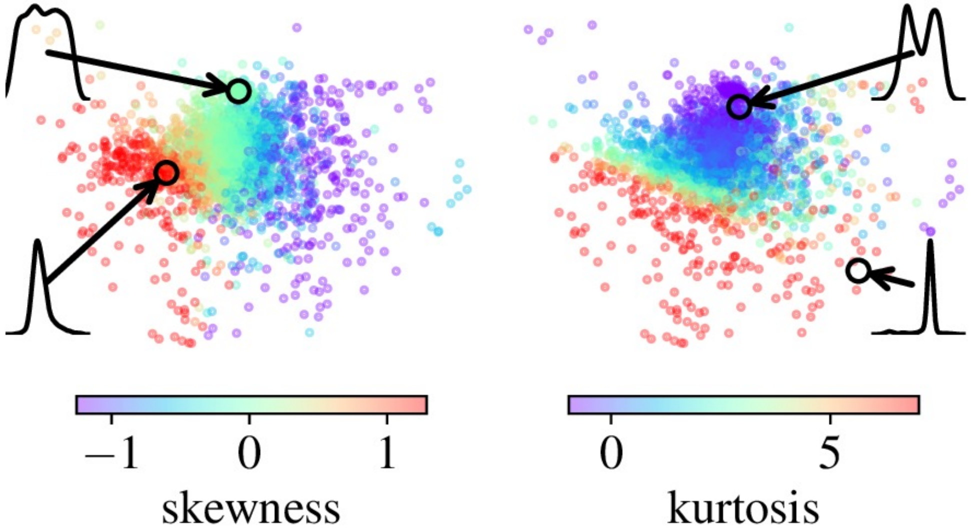
	Pretraining			***Inference***		
	# of samples	# of features	# of classes	# of samples	# of features	# of classes
TabPFN	1,024	100	10	3,000	100	10
TabPFNv2	2,048	160	10	10,000	500	10
TabICL	60,000	100	10	100,000	500	Any

Distribution-aware Column-wise Embedding

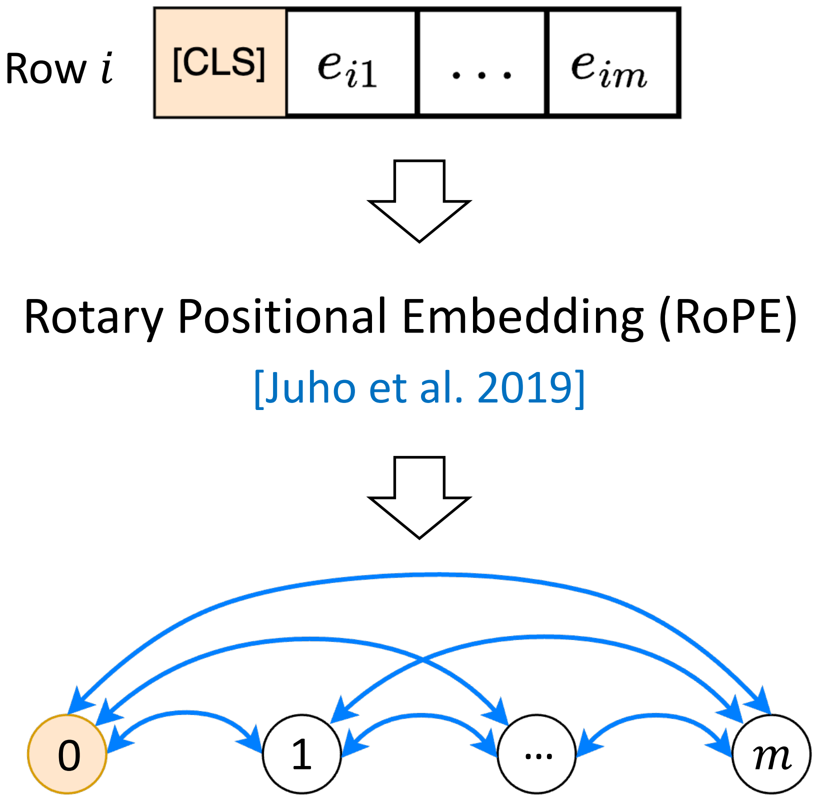
Set Transformer
(Linear attention)

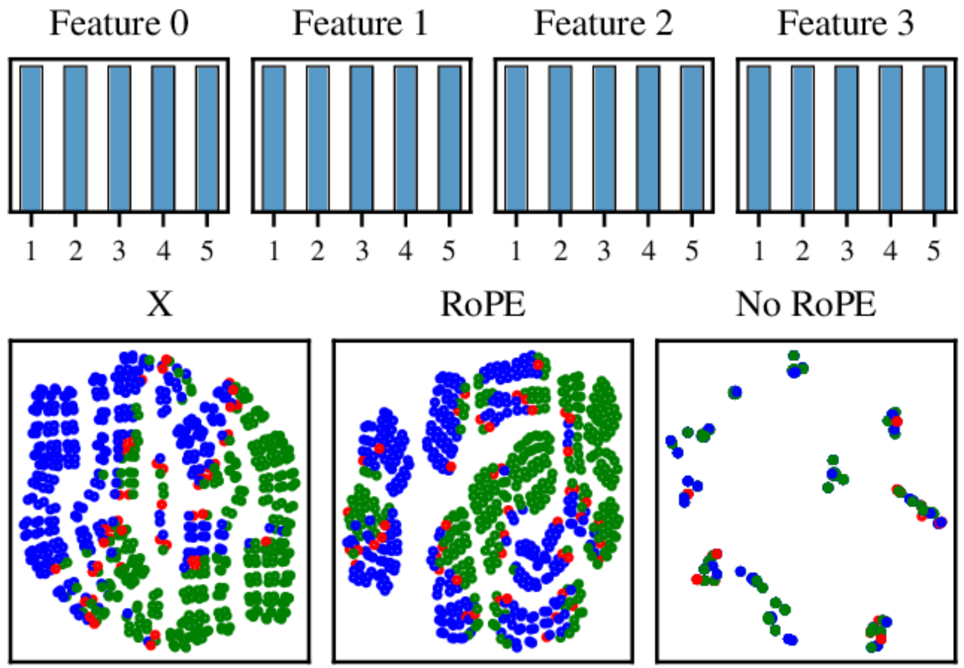
[Juho et al. 2019]





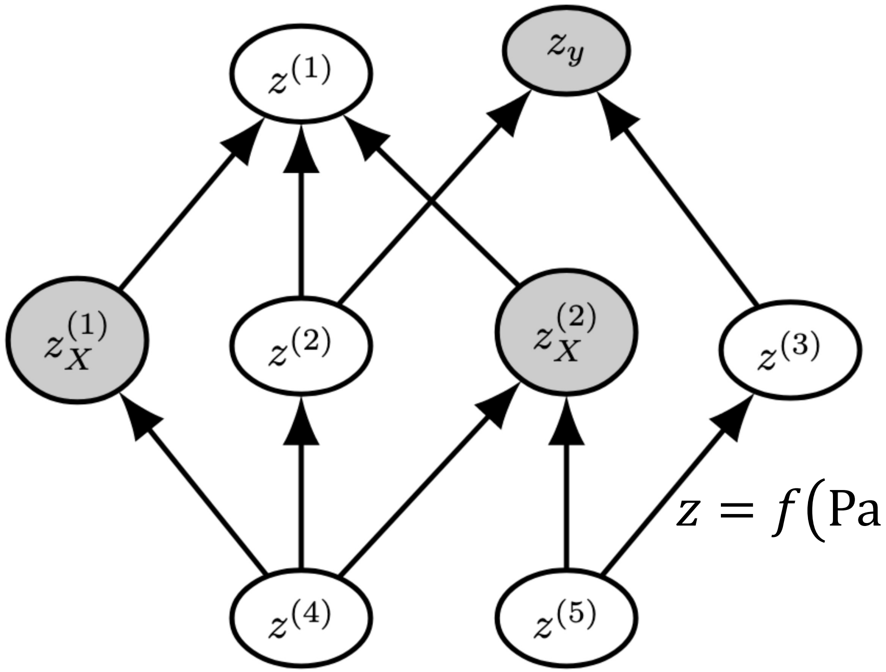
Learned embeddings encode statistical properties of feature distributions.





RoPE alleviates the representation collapse.

Synthetic Prior Datasets

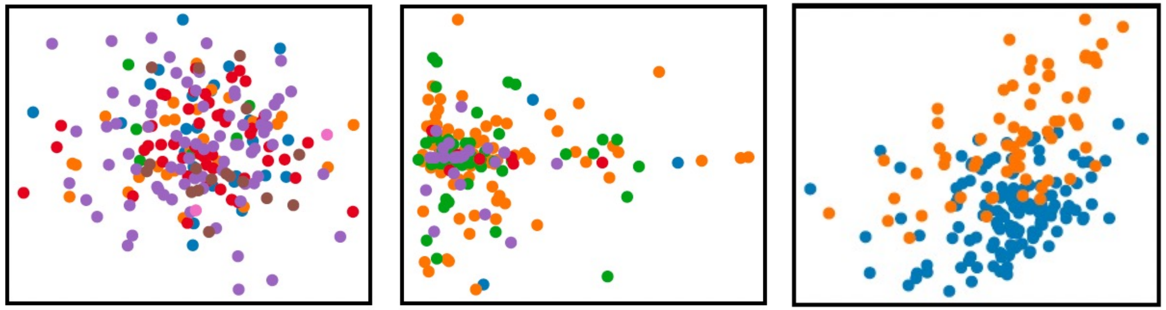


$z = f\big(\mathrm{Pa}(z)\big) + \varepsilon$

Structural causal models

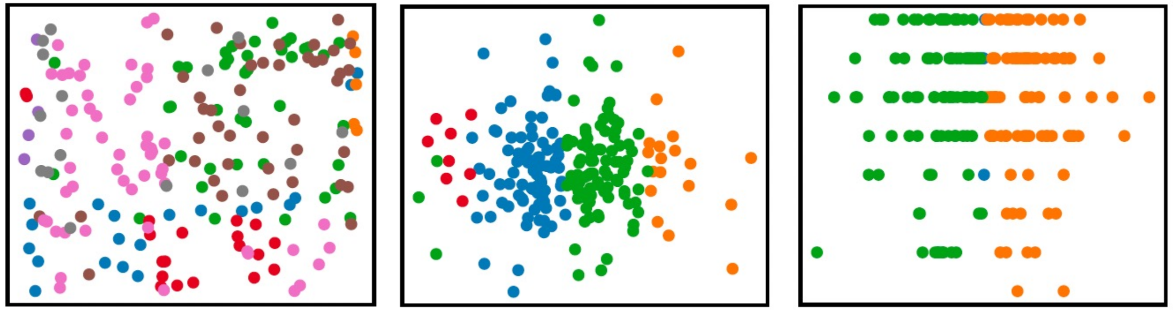
[Hollmann et al. 2022]

MLP-based SCM (70%) f : Activation ∘ Linear



The weights and biases of the linear layer are randomly sampled.

Tree-based SCM (30%) f : XGBoost [Chen et al. 2016]

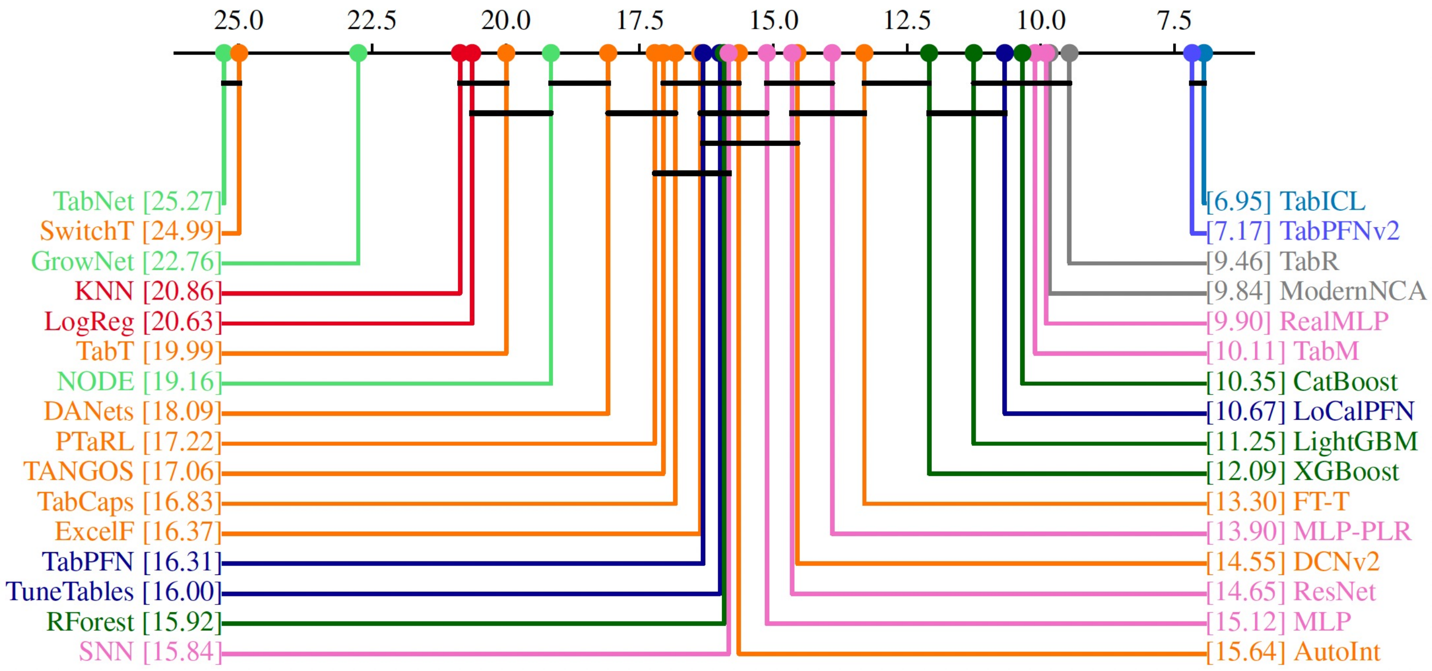


XGBoost is trained on fake targets drawn from Gaussian noise.

Comparison between TabPFNv2 and TabICL

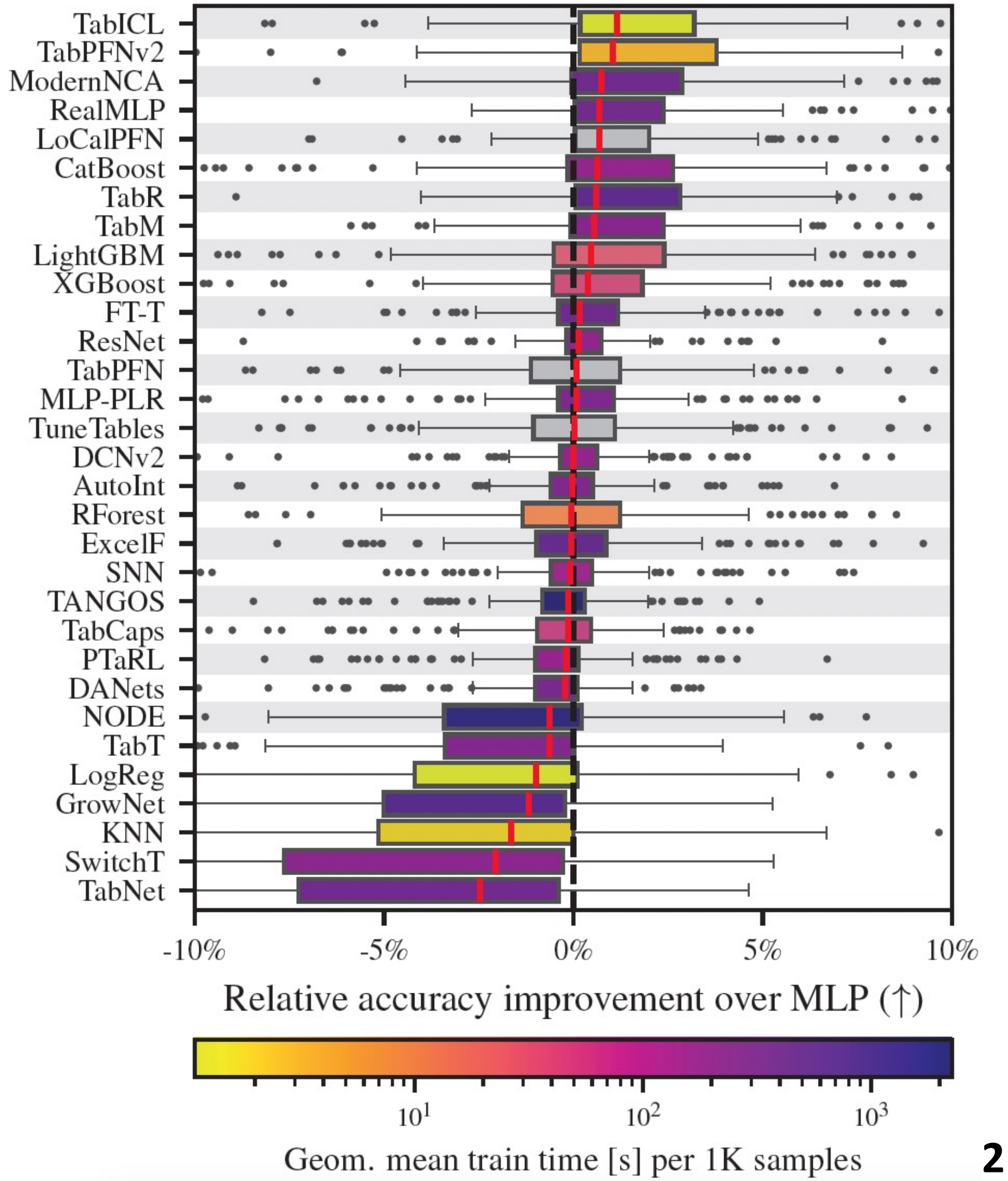
	Aspect	TabPFNv2	TabICL
Architecture	Attention	Alternating column / row attentions	Column -> Row -> ICL
	Collapse Issue	Feature grouping and random feature vectors	RoPE
	Label Fusion	Early (input layer)	Late (for ICL)
Pre-training	SCM	Growing network with redirection (Code not open-source)	Layered structure
	# of datasets	82 million	130 million
	Curriculum learning	❎ # of samples $\leq$ 2,048	✅ # of samples 1K -> 60K
Scalability	# of samples	$\leq$ 10K	100K (only 5GB GPU)
	# of classes	$\leq$ 10	Any (Hierarchical classification)

Performance of TabICL on the TALENT Benchmark

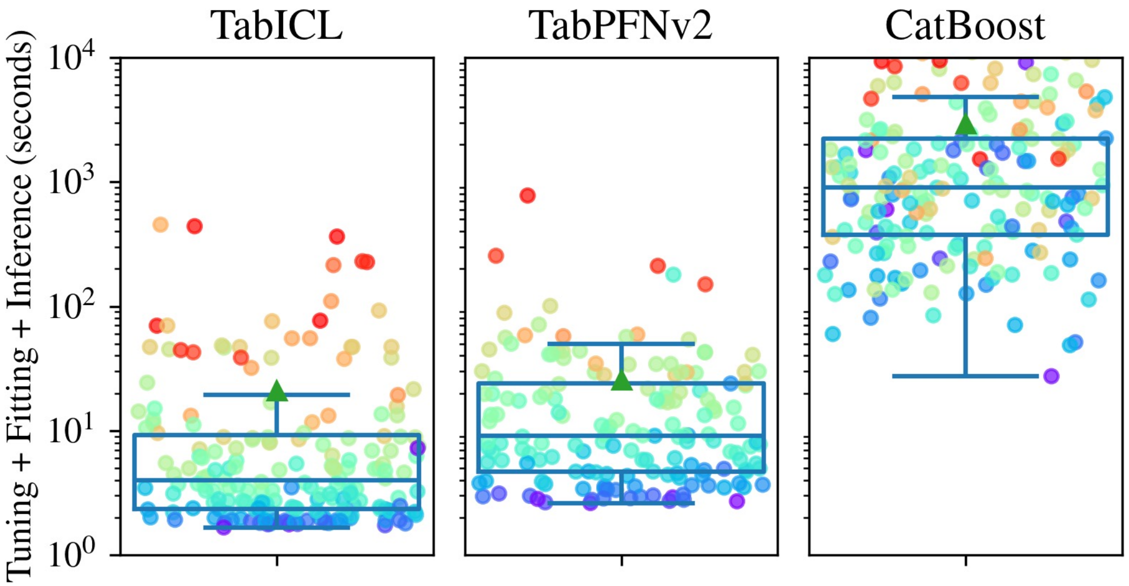


Average accuracy ranking over 171 classification datasets (≤ 10 classes) from the TALENT benchmark

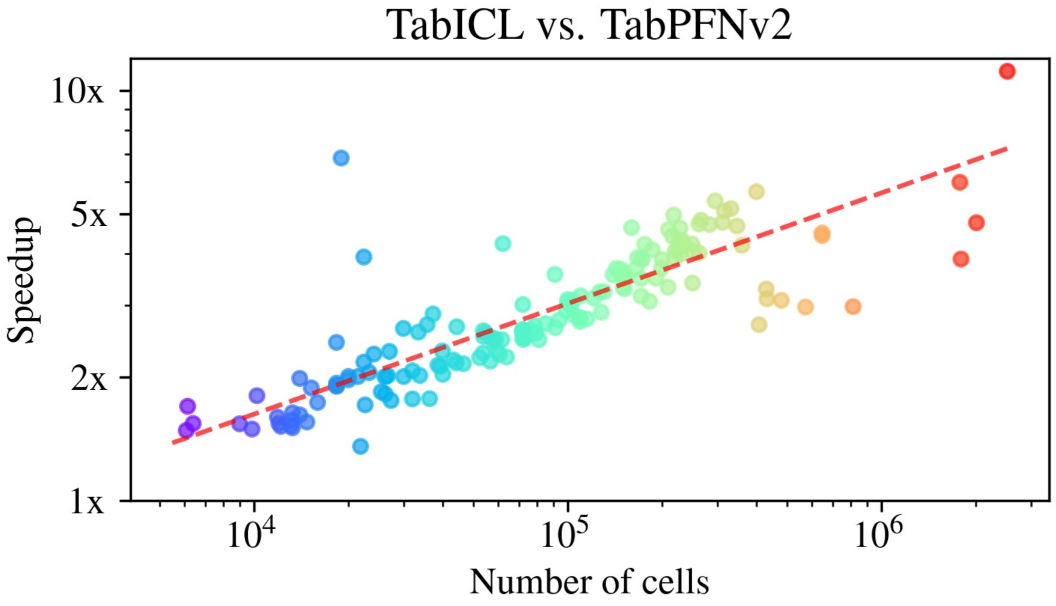
[Ye et al. 2024]



TabICL is faster

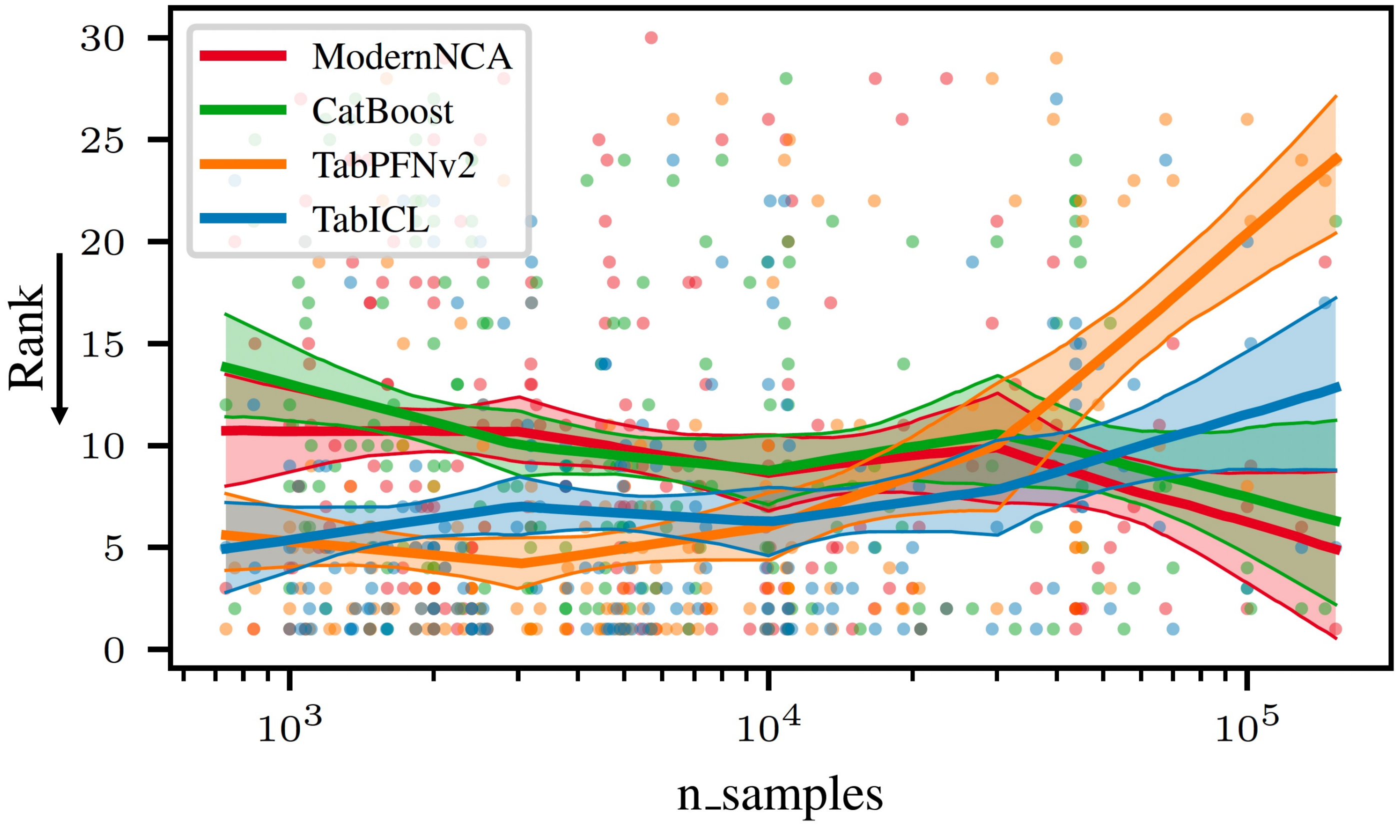


Time comparison per dataset



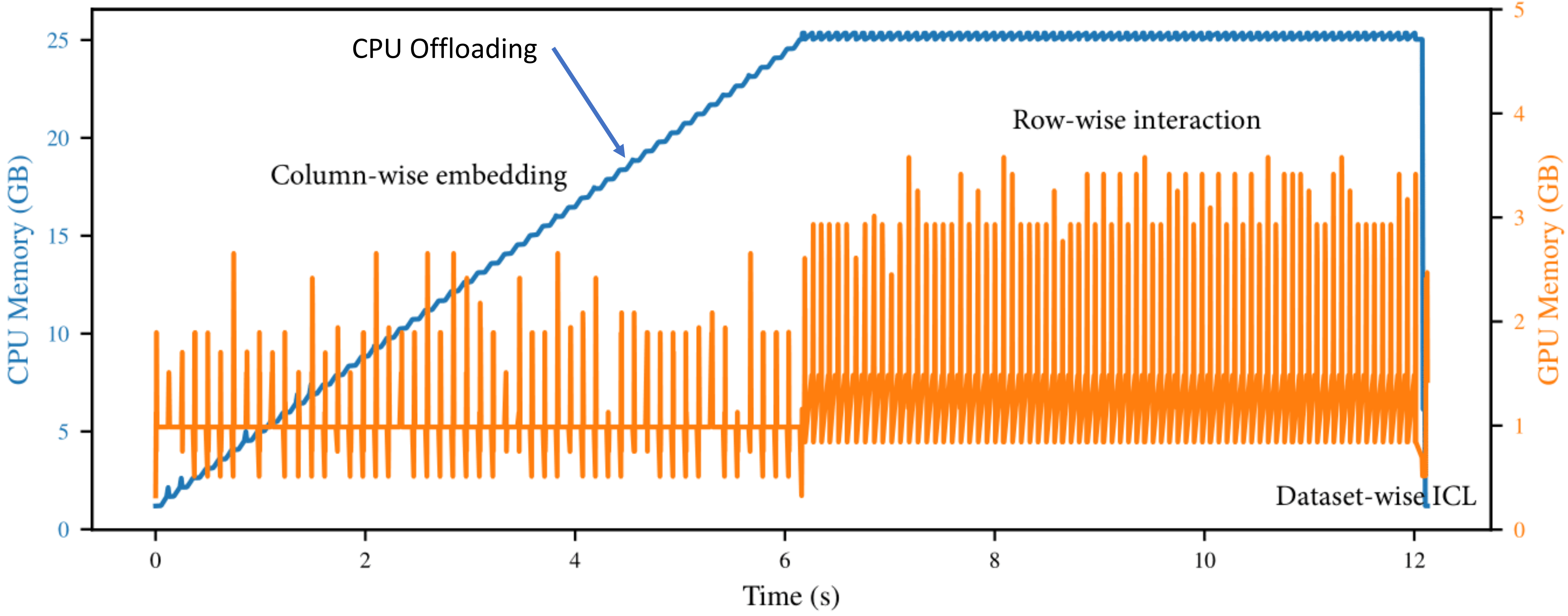
Speedup of TabICL vs. TabPFNv2

TabICL scales effectively with data size



Model rankings vs. sample size

TabICL enables memory-efficient inference



CPU and GPU memory during inference for a dataset with 100K samples and 500 features (**< 5GB GPU memory**)

- Tabular foundation models can be scaled to an order of magnitude larger data!
  - Expressive yet computationally efficient architecture
  - Large-scale pre-training via curriculum learning by gradually increasing data size
  - Memory-efficient inference through CPU offloading, adaptive batching, etc.

- Faster and better than TabPFNv2 despite
  - Simpler preprocessing
  - Simpler prior (no random graphs, no categorical vectors/discretization)

- Open-source everything : https://github.com/soda-inria/tabicl

Thanks for your attention !