TabCBM: Concept-based Interpretable Neural Networks for Tabular Data

Paper + Code!

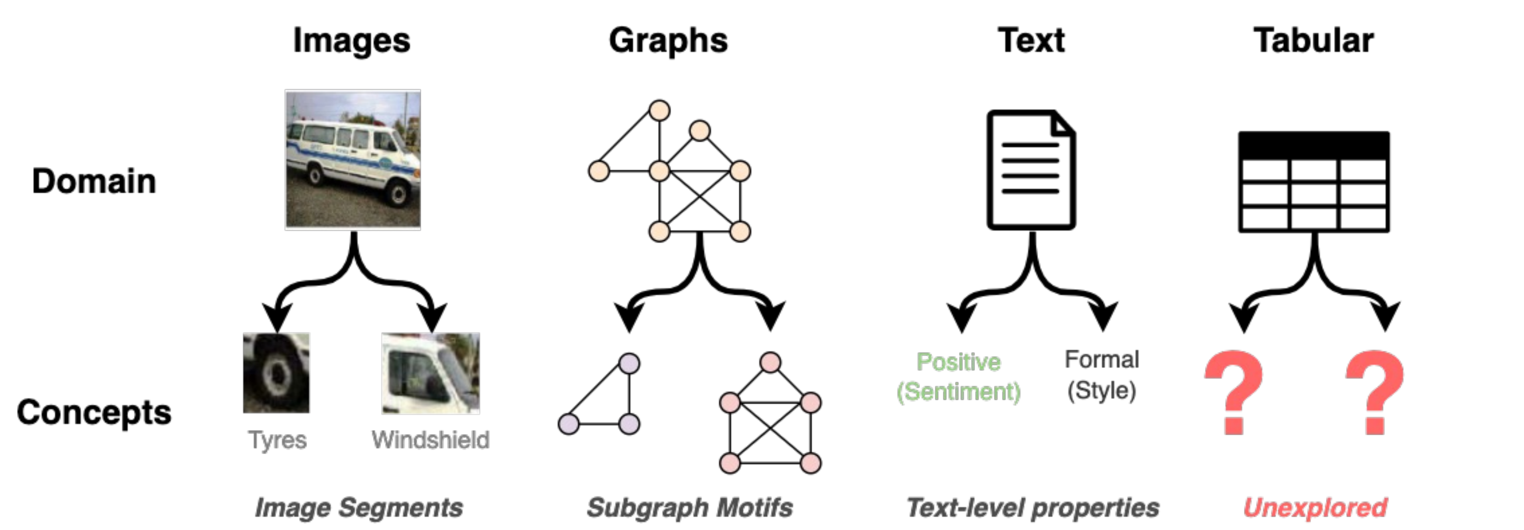
Mateo Espinosa Zarlenga, Zohreh Shams, Michael E. Nelson, Been Kim*, Mateja Jamnik*

UNIVERSITY OF CAMBRIDGE · Google DeepMind · GATES Cambridge

Paper + Code!

Research Gap: How do tabular tasks fit within concept-based interpretable frameworks?

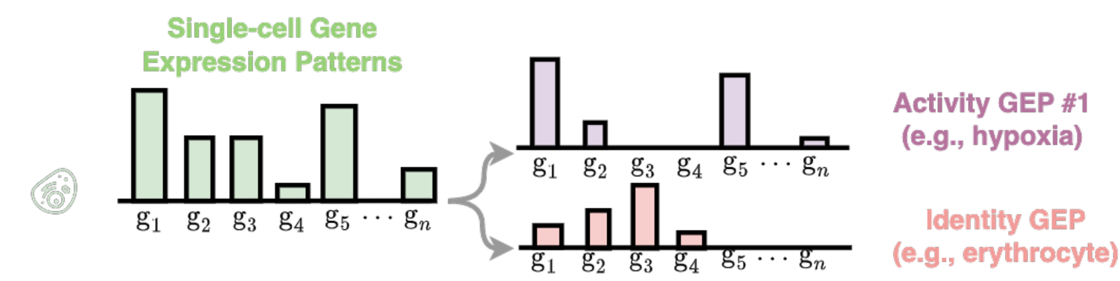
- Recent work in explainable artificial intelligence (XAI) [1-4] has proposed interpretable neural networks that explain predictions via high-level "concepts".
- However, previous works in this field have been uniquely focused on image [2], graph-structured [3], and text [4] tasks, **leaving crucial tabular tasks, such as clinical and genomics tasks, outside of the scope of these methods**.



- Hence, in this work we explore (1) **what a concept entails in a tabular task** and (2) **how we can construct concept-interpretable models** without sacrificing the performance observed in simpler state-of-the-art tabular methods (e.g., GBMs).

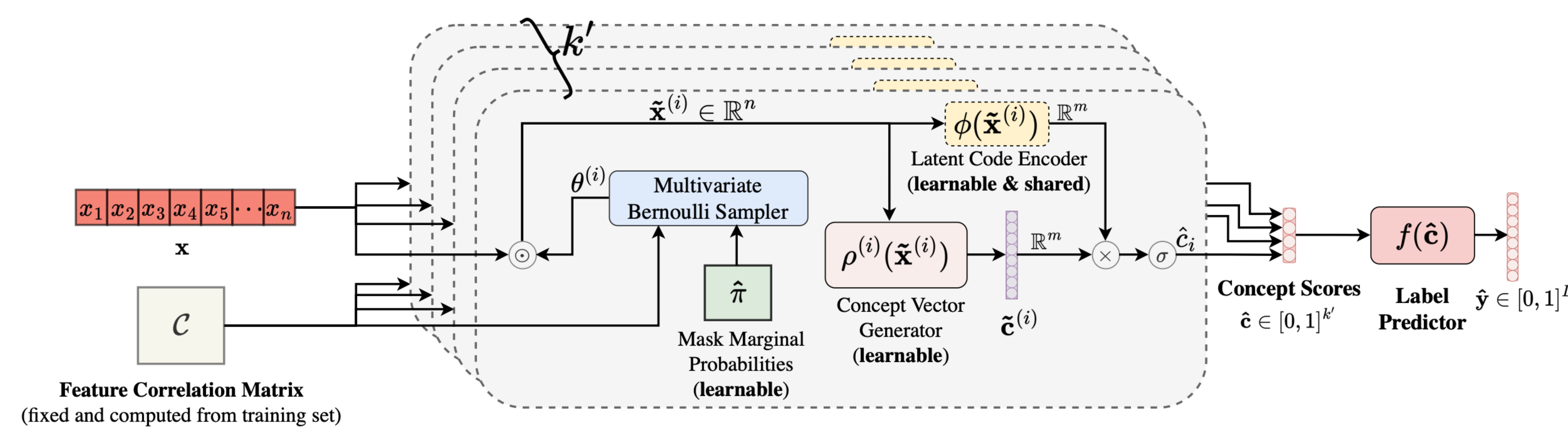
Key idea: Feature subsets as tabular concepts

Given a task on n input features, we define a tabular concept as a fixed **group of highly correlated features** $\boldsymbol{\pi} \in [0,1]^n$ that form the input to a scoring function representing a **"meta feature"** $s: \mathbb{R}^{\sum \pi_i} \to \{0,1\}$.



Tabular Concept Bottleneck Model (TabCBM)

We discover concepts via a **differentiable feature selection** mechanism that learns k' pairs $\{(\hat{\boldsymbol{\pi}}^{(i)}, s^{(i)})\}_{i=1}^{k'}$ of subsets of features $\hat{\boldsymbol{\pi}}^{(i)}$ and scoring functions $s^{(i)}$ from which a **bottleneck of concept scores** $\hat{\mathbf{c}} \in [0,1]^k$ can be used to predict a downstream task.



Training: How do we learn meaningful concepts?

We include regularisers that encourage:

1. **Completeness** → discovered concept scores $\hat{\mathbf{c}}$ should predict a task of interest.

$$\mathcal{L}_\text{task}(f(\hat{\mathbf{c}}), y)$$

2. **Coherency**→ Similar samples should lead to a similar set of concept scores.

$$\mathcal{L}_\text{co}(\mathbf{x}_1, \cdots, \mathbf{x}_N) := -\frac{1}{Nt} \sum_{\mathbf{x}_i \in \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}} \sum_{\phi(\mathbf{x}_j) \in \Psi_t(\phi(\mathbf{x}_i))} \frac{\hat{\mathbf{c}}(\mathbf{x}_i)^T \hat{\mathbf{c}}(\mathbf{x}_j)}{||\hat{\mathbf{c}}(\mathbf{x}_i)||\,||\hat{\mathbf{c}}(\mathbf{x}_j)||}$$

3. **Diversity** → different scoring functions and masks represent different concepts.

$$\mathcal{L}_\text{div}(\mathbf{x}_1, \cdots, \mathbf{x}_N) := \frac{1}{Nk'(k'-1)} \sum_{\mathbf{x} \in \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}} \sum_{i=1}^{k'} \sum_{\substack{j=1 \\ j \neq i}}^{k'} \frac{\rho_j(\tilde{\mathbf{x}}^{(j)})^T \rho_i(\tilde{\mathbf{x}}^{(i)})}{||\rho_j(\tilde{\mathbf{x}}^{(j)})||\,||\rho_i(\tilde{\mathbf{x}}^{(i)})||}$$

4. **Specificity** → concepts should be a function of only a handful of input features.

$$\mathcal{L}_\text{spec}(\hat{\boldsymbol{\pi}}^{(1)}, \cdots, \hat{\boldsymbol{\pi}}^{(k')}) := \frac{1}{k'n} \sum_{i=1}^{k'} ||\hat{\boldsymbol{\pi}}^{(i)}||_1$$

Furthermore, as in traditional concept bottleneck models (CBMs) [1], **we can include supervision for known concepts** when we have train-time concept labels.

References

[1] Koh, Pang Wei, et al. "Concept bottleneck models." *International Conference on Machine Learning*. PMLR, 2020.

[2] Ghorbani, Amirata, et al. "Towards automatic concept-based explanations." *Advances in neural information processing systems* 32 (2019).

[3] Magister, Lucie Charlotte, et al. "GCExplainer: Human-in-the-loop concept-based explanations for graph neural networks." *arXiv preprint arXiv:2107.11889* (2021).

[4] Yeh, Chih-Kuan, et al. "On completeness-aware concept-based explanations in deep neural networks." *Advances in neural information processing systems* 33 (2020): 20554-20565.

Main Results

Key Finding #1: Interpretability without sacrificing performance

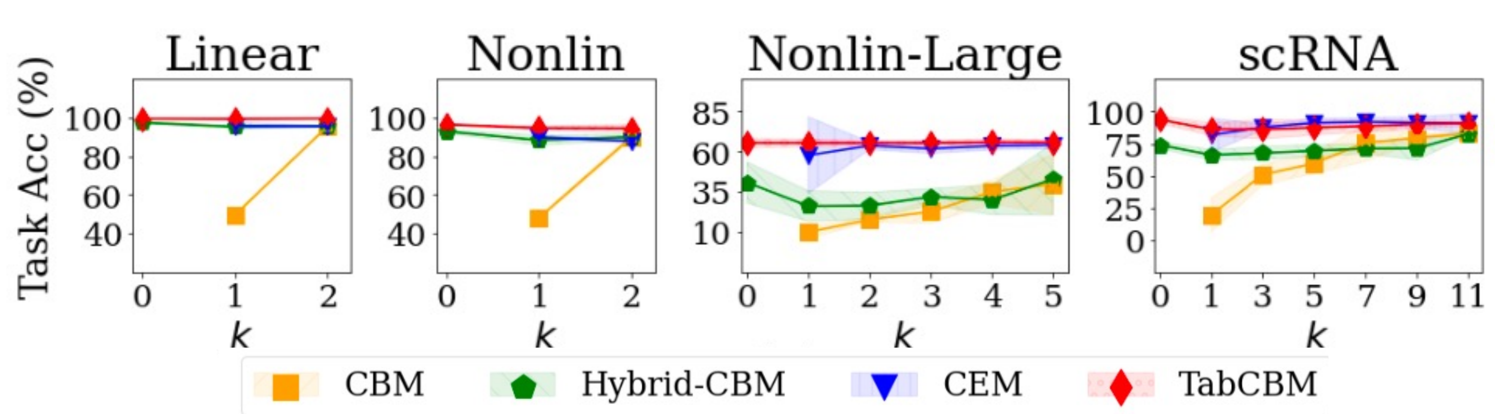


Figure 1: Task accuracy (%) of concept-interpretable methods across synthetic tabular tasks with known ground truth concepts. We show the accuracy as we vary the number of training concepts k.

Dataset	TabCBM (ours)	SENN	CCD (recon)	MLP	TabNet	TabTransformer	XGBoost	LightGBM
Synth-Linear	**99.84 ± 0.06**	98.15 ± 0.2	96.47 ± 1.3	97.57 ± 0.37	97.57 ± 0.37	82.91 ± 0.55	96.43	96.8
Synth-Nonlin	**98.36 ± 0.15**	89.14 ± 0.71	85.99 ± 2.28	87.65 ± 0.98	91.57 ± 0.48	81.07 ± 0.83	88.43	89.33
Synth-Nonlin-Large	**62.78 ± 1.13**	49.78 ± 2.08	51.64 ± 1.71	40.73 ± 6.42	51.01 ± 2.57	54.63 ± 1.17	22.48 ± 0.48	23.58 ± 0.78
Synth-scRNA	**93.66 ± 1.41**	78.32 ± 3.03	68.83 ± 1.73	73.87 ± 1.43	90.66 ± 1.10	87.29 ± 0.68	90.44 ± 1.06	89.96 ± 1.57
Higgs (without high-level)	**80.42 ± 0.3**	70.61 ± 0.12	77.84 ± 0.08	79.90 ± 0.15	79.44 ± 0.16	74.94 ± 0.21	68.85 ± 0.02	68.87 ± 0.06
Higgs (with high-level)	**78.62 ± 0.12**	73.53 ± 0.71	77.92 ± 0.09	78.44 ± 0.02	78.12 ± 0.05	74.22 ± 0.42	75.33 ± 0.04	75.33 ± 0.03
PBMC	**93.35 ± 0.16**	92.24 ± 0.23	93.14 ± 0.30	91.66 ± 1.95	92.74 ± 0.46	91.01 ± 0.33	93.09 ± 0.29	93.01 ± 0.24
FICO	72.08 ± 0.42	66.78 ± 0.69	65.46 ± 4.91	67.98 ± 1.36	71.20 ± 0.87	65.66 ± 0.85	72.33 ± 0.44	**72.63 ± 0.12**

Table 1: Task accuracy (%) of competing methods across tabular tasks *without* ground truth concept labels at train time.

Key Finding #2: TabCBM discovers tabular concepts aligned with expert-annotated concepts

	CAS (*coherence*)	MIG (*diversity*)	R^4 (*coherence & diversity*)	Dis (*diversity*)	Compl (*completeness*)
TabCBM (ours)	**87.55 ± 14.07** ($\bar{r}$ = 1.5)	**57.71 ± 26.27** ($\bar{r}$ = 1.5)	**78.36 ± 17.65** ($\bar{r}$ = 1.5)	**69.83 ± 23.65** ($\bar{r}$ = 1.5)	**70.44 ± 22.81** ($\bar{r}$ = 1.5)
SENN	60.11 ± 6.26 ($\bar{r}$ = 2.75)	9.92 ± 5.68 ($\bar{r}$ = 3.5)	30.83 ± 17.38 ($\bar{r}$ = 3.5)	21.49 ± 6.51 ($\bar{r}$ = 3.5)	29.56 ± 7.30 ($\bar{r}$ = 3.75)
CCD	52.86 ± 20.82 ($\bar{r}$ = 3)	29.57 ± 5.86 ($\bar{r}$ = 2)	65.79 ± 10.49 ($\bar{r}$ = 2)	39.66 ± 5.89 ($\bar{r}$ = 2)	41.04 ± 6.93 ($\bar{r}$ = 2.25)
PCA	57.54 ± 12.89 ($\bar{r}$ = 2.75)	9.48 ± 5.73 ($\bar{r}$ = 3)	19.59 ± 28.18 ($\bar{r}$ = 3)	24.15 ± 16.9 ($\bar{r}$ = 3)	36.17 ± 15.86 ($\bar{r}$ = 2.25)

Table 2: Mean concept representation quality metrics (%) measured across several synthetic datasets with ground-truth concept annotations (higher values are better).

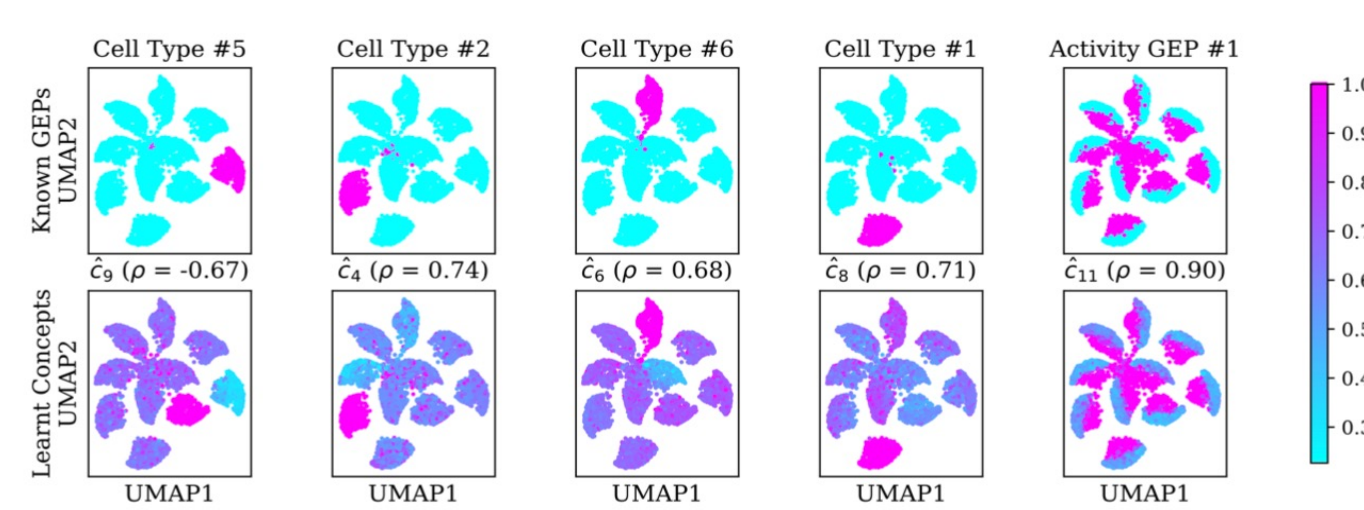


Figure 2: Five known Gene Expression Programs (GEPs) in a synthetic scRNA task together with TabCBM's discovered concept with the highest absolute correlation with each GEP.

Key Finding #3: Performance can be boosted via human-in-the-loop concept interventions

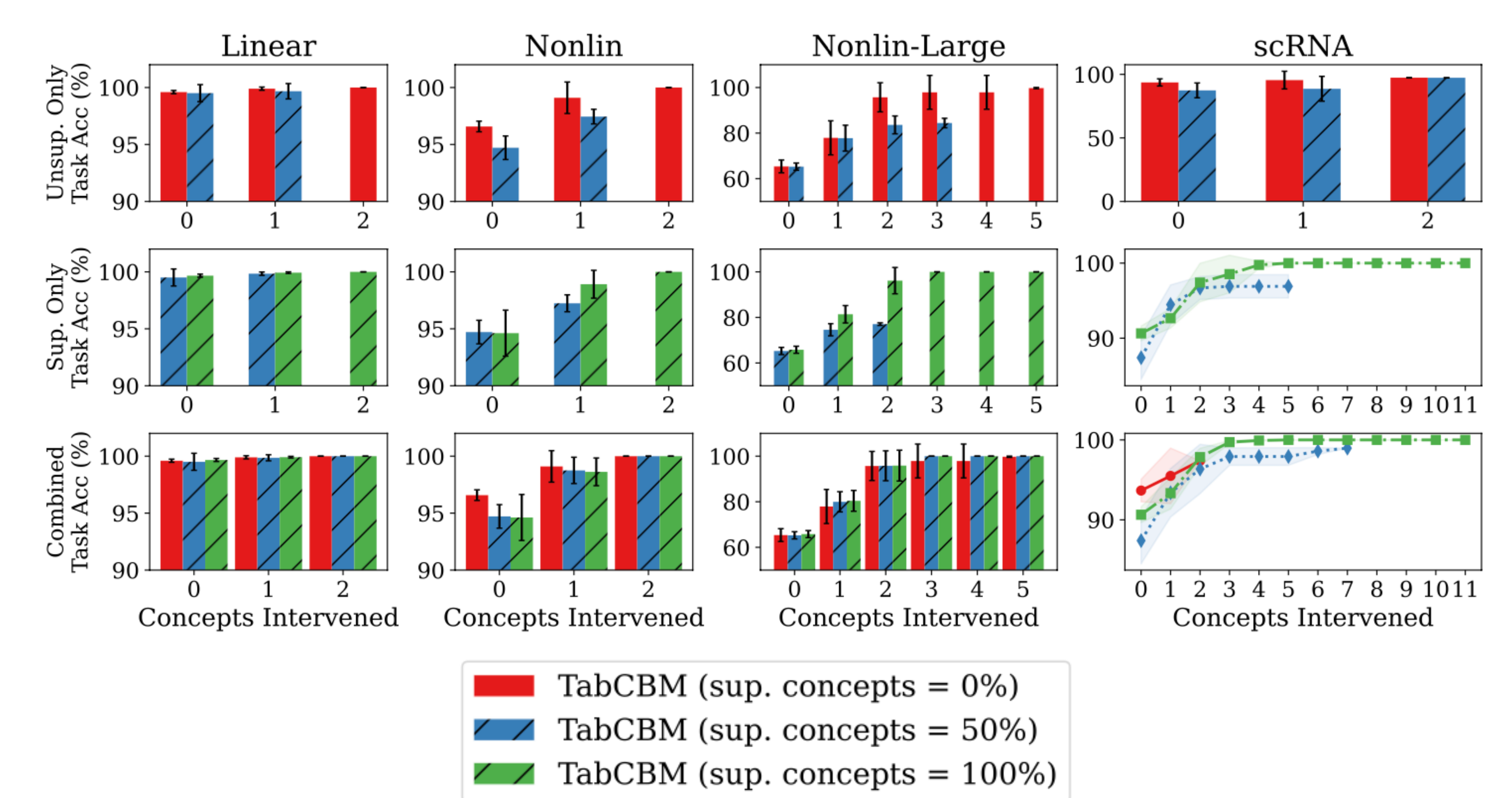


Figure 3: TabCBM task accuracy after intervening on a varying number of concepts (x-axis), across tasks (columns), and varying whether we intervene only on supervised concepts (rows).