In Search of Grandmother Cells: Tracing Interpretable Neurons in Tabular Representations

Ricardo Knauer
KI-Werkstatt
University of Applied Sciences Berlin
Berlin, Germany

Erik Rodner
KI-Werkstatt
University of Applied Sciences Berlin
Berlin, Germany

Abstract

Foundation models are powerful yet often opaque in their decision-making. A topic of continued interest in both neuroscience and artificial intelligence is whether some neurons behave like "grandmother cells", *i.e.*, neurons that are inherently interpretable because they exclusively respond to single concepts. In this work, we propose two information-theoretic measures that quantify the neuronal saliency and selectivity for single concepts. We apply these metrics to the representations of TabPFN, a tabular foundation model, and perform a simple search across neuron-concept pairs to find the most salient and selective pair. Our analysis provides the first evidence that some neurons in such models show moderate, statistically significant saliency and selectivity for high-level concepts. These findings suggest that interpretable neurons can emerge naturally and that they can, in some cases, be identified without resorting to more complex interpretability techniques.

1 Introduction

The rise of foundation models has unlocked powerful capabilities, but their decision-making processes often remain opaque (Bommasani et al., 2022; Longo et al., 2024). A recurring theme in the fields of neuroscience (Gross, 2002; Quiroga et al., 2005) and artificial intelligence (Arora et al., 2018; Olah et al., 2020) is whether some neurons act like "grandmother cells", *i.e.*, neurons that exclusively respond to single high-level features or concepts. Understanding the nature of neuronal representations can facilitate their interpretation and foster trust among users, especially in high-stake domains such as the clinical setting (Adler et al., 2022; European Parliament and Council of the European Union, 2024). In recent years, many advanced techniques have been developed within the subfield of mechanistic interpretability to gain deeper insights into the inner workings of foundation models, *e.g.*, by identifying interpretable units of computation (Ferrando et al., 2024; Sharkey et al., 2025). In our work, we examine whether we actually always need sophisticated methods to find interpretable units.

Our contributions are as follows:

- We propose **two information-theoretic measures of neuronal saliency and selectivity for single concepts** (Hewitt and Liang, 2019; Kandel et al., 2021; Simonyan et al., 2014). We use hypothesis testing to find neurons that are significantly salient and selective for single concepts and analytically derive the null distributions for our metrics (Sect. 3).
- We apply our metrics to the representations of TabPFN (Hollmann et al., 2025), a tabular foundation model, and offer the first evidence that **some neurons in tabular foundation models exhibit moderate, statistically significant saliency and selectivity for high-level concepts** (Sect. 4). Our findings suggest that interpretable neurons can emerge naturally and that they can sometimes be uncovered without relying on more complex interpretability techniques.

EurIPS 2025 Workshop on AI for Tabular Data

2 Related work

Whether individual neurons correspond to single concepts, *i.e.*, behave like "grandmother cells", remains a fundamental question in both neuroscience (Gross, 2002; Quiroga et al., 2005) and artificial intelligence (Arora et al., 2018; Olah et al., 2020). Formally, a neural network cannot represent concepts using orthogonal bases if the number of independent concepts is greater than the number of neurons in the network. As a result, concepts are often encoded by multiple neurons (distributed representation) and neurons often encode multiple concepts (superposition) (Arora et al., 2018; Olah et al., 2020). Although there are interpretability techniques to localize and disentangle such representations, *e.g.*, sparse probes (Bertsimas et al., 2021; Gurnee et al., 2023) and sparse autoencoders (Gao et al., 2025; Lieberum et al., 2024; Templeton et al., 2024), they usually require training auxiliary models and incur high computational costs. This raises the question whether simpler methods could sometimes suffice to identify interpretable neurons.

In the next section, we propose to quantify the neuronal saliency and selectivity for single concepts (Hewitt and Liang, 2019; Kandel et al., 2021; Simonyan et al., 2014) using information theory (Dayan and Abbott, 2005; Li et al., 2024; Shannon, 1948), without requiring the training of auxiliary models.

3 Quantifying neuronal interpretability

In this section, we define our two information-theoretic measures of neuronal saliency and selectivity for single concepts, and analytically derive their null distributions for hypothesis testing.

3.1 Metric definitions

Let $\mathbf{A} \in \mathbb{R}^{M \times N}$ be the neuronal activations for $M \in \mathbb{N}_+$ samples and $N \in \mathbb{N}_+$ neurons. The network structure is not relevant for the subsequent definitions and therefore abstracted away. Moreover, let $\mathbf{B} \in \{0,1\}^{M \times C}$ represent the concept labels for the M samples and $C \in \mathbb{N}_+$ high-level, interpretable features, *i.e.*, concepts. We measure the association between the activations $\mathbf{a}_i \in \mathbb{R}^M$ of neuron i and the concept labels $\mathbf{b}_j \in \{0,1\}^M$ of concept j with the empirical mutual information $\hat{I}(\mathbf{a}_i, \mathbf{b}_j) \in \mathbb{R}_{+\cup\{0\}}$ (Dayan and Abbott, 2005; Li et al., 2024; Shannon, 1948). We further denote by

$$\hat{p}_{tail}(\mathbf{a}_i, \mathbf{b}_j) = \frac{1}{N} \sum_{n=1}^{N} [\hat{I}(\mathbf{a}_n, \mathbf{b}_j) \geq \hat{I}(\mathbf{a}_i, \mathbf{b}_j)], \quad \frac{1}{N} \leq \hat{p}_{tail}(\mathbf{a}_i, \mathbf{b}_j) \leq 1 \tag{1}$$

the empirical upper-tail probability of $\hat{I}(\mathbf{a}_i, \mathbf{b}_j)$, *i.e.*, the fraction of neurons whose empirical mutual information to concept j is greater than or equal to that of neuron i, where $[\cdot]$ are the Iverson brackets.

We quantify the saliency (Kandel et al., 2021; Simonyan et al., 2014) for neuron i with respect to concept j using the information content or surprisal of observing $\hat{p}_{tail}(\mathbf{a}_i, \mathbf{b}_j)$. This captures the intuition that unexpected observations indicate more pronounced neuron–concept relationships:

$$\text{surprisal}(\mathbf{a}_i, \mathbf{b}_j) = -\log(\hat{p}_{tail}(\mathbf{a}_i, \mathbf{b}_j)), \quad 0 \leq \text{surprisal}(\mathbf{a}_i, \mathbf{b}_j) \leq \log(N) \quad . \tag{2}$$

Then, we define the selectivity (Hewitt and Liang, 2019; Kandel et al., 2021) as:

$$\text{selectivity}(\mathbf{a}_i, \mathbf{b}_j) = \frac{\text{surprisal}(\mathbf{a}_i, \mathbf{b}_j)}{\sum_{c=1}^{C} \text{surprisal}(\mathbf{a}_i, \mathbf{b}_c)}, \quad 0 \leq \text{selectivity}(\mathbf{a}_i, \mathbf{b}_j) \leq 1 \quad . \tag{3}$$

Please note that smaller concept set sizes tend to yield higher selectivity values due to the normalization, and that the selectivity only accounts for concepts $c \in \{1, ..., C\}$. In the following, we derive the upper-tail p-values for $\text{surprisal}(\mathbf{a}_i, \mathbf{b}_j)$ and $\text{selectivity}(\mathbf{a}_i, \mathbf{b}_j)$ in isolation and in combination.

3.2 Significance testing

Our null hypothesis H_0 is that neuron-concept associations are exchangeable, *i.e.*, no neuron is preferentially associated with any concept and no concept is preferentially associated with any neuron. We assume that the mutual information scores are independent and identically distributed (i.i.d.) across neurons for concept j and that the surprisals are i.i.d. across concepts for neuron i. As $N \to \infty$, this implies that $\hat{p}_{tail}(\cdot, \mathbf{b}_j) \Rightarrow \text{Uniform}(0,1)$ and $\text{surprisal}(\cdot, \mathbf{b}_j) \Rightarrow \text{Exponential}(1) = \text{Gamma}(1,1)$.

Normalizing C i.i.d. Gamma(1, 1) random variables yields a flat Dirichlet$(1, ..., 1)$ distribution with Beta$(1, C-1)$ marginals, so that the selectivity for neuron i with respect to each concept $j \in \{1, ..., C\}$ is Beta$(1, C-1)$-distributed.

The upper-tail p-value of observing surprisal$(\mathbf{a}_i, \mathbf{b}_j)$ is:

$$P(\text{surprisal}(\cdot, \mathbf{b}_j) \geq \text{surprisal}(\mathbf{a}_i, \mathbf{b}_j) \mid H_0) = \hat{p}_{tail}(\mathbf{a}_i, \mathbf{b}_j) \quad . \tag{4}$$

The upper-tail p-value of observing selectivity$(\mathbf{a}_i, \mathbf{b}_j)$ is (Olver et al., 2010):

$$P(\text{selectivity}(\mathbf{a}_i, \cdot) \geq \text{selectivity}(\mathbf{a}_i, \mathbf{b}_j) \mid H_0) = (1 - \text{selectivity}(\mathbf{a}_i, \mathbf{b}_j))^{C-1} \quad . \tag{5}$$

To enforce a conservative family-wise error rate control, we combine both p-values with a Bonferroni correction that adjusts the minimum p-value for the number of tests (Nikolitsa et al., 2025):

$$p_{comb}(\mathbf{a}_i, \mathbf{b}_j) = \min(2NC \cdot \min(\hat{p}_{tail}(\mathbf{a}_i, \mathbf{b}_j), (1 - \text{selectivity}(\mathbf{a}_i, \mathbf{b}_j))^{C-1}), 1) \quad . \tag{6}$$

In the next section, we use our definitions of surprisal$(\mathbf{a}_i, \mathbf{b}_j)$ (Eq. (2)) and selectivity$(\mathbf{a}_i, \mathbf{b}_j)$ (Eq. (3)) as well as $p_{comb}(\mathbf{a}_i, \mathbf{b}_j)$ (Eq. (6)) to empirically test whether we can use our metrics to find interpretable neurons in foundation models.

4 Experiments

In this section, we use the surprisal and selectivity (Sect. 3.1) as well as their combined p-value (Sect. 3.2) to evaluate whether we can use our measures to uncover interpretable neurons in foundation models. To this end, we apply them to the representations of TabPFN (Hollmann et al., 2025), a tabular foundation model, and provide the first evidence that some neurons in such models show moderate, statistically significant saliency and selectivity for high-level concepts. These findings suggest that interpretable neurons can emerge naturally and that, in some cases, they can be identified without relying on more complex interpretability techniques.

4.1 Experimental setup

We focused our experiments on a setting with clearly defined concept labels. Specifically, we treated diagnostic codes from the International Classification of Diseases (ICD) as concepts organized at three hierarchical levels (low, mid, and high) and considered them for four tabular prediction tasks at emergency department (ED) triage: inhospital mortality, intensive care unit (ICU) transfer within 12h, critical outcome, and hospitalization prediction (Xie et al., 2022). Please refer to Appendix A for details on the datasets and preprocessing.

We employed TabPFN 2.0.9 (Hollmann et al., 2025) as our foundation model due to its strong reported performance (Erickson et al., 2025). We used TabPFN without tuning or ensembling and evaluated its discriminative performance with the test set area under the receiver operating characteristic curve (AUC). We then extracted 192-dimensional embeddings from the 12 encoder layers, computed the surprisal and selectivity for all neuron-concept pairs, and treated the knee point on the surprisal-selectivity Pareto front as our most salient and selective neuron-concept pair. To identify the knee point, we maximized the sum of the min-max scaled surprisal and selectivity scores. We additionally computed $p_{comb}(\mathbf{a}_i, \mathbf{b}_j)$ (Eq. (6)) for the knee point and selected the top-3 input features for the knee point activations based on their mutual information when the concept was present. As baselines, we employed sparse probes based on SHAP values (Covert et al., 2021; Lundberg and Lee, 2017) and optimal probing (Bertsimas et al., 2021; Gurnee et al., 2023). Please refer to Appendix B for details.

4.2 Experimental results

TabPFN's test set AUC ranged from acceptable (0.79) to excellent (0.88), matching the best machine learning baselines in Xie et al. (2022). Most neurons in TabPFN were neither salient nor selective for specific ICD codes (Fig. 1 and 2 in Appendix C). Nevertheless, some neuron-concept pairs could reach maximum surprisal values of $\log(N) = 7.74$ or moderate selectivity values of up to 0.73. Sparse probes based on SHAP values and optimal probing were Pareto-dominated by our search procedure (Fig. 1 and 2 in Appendix C).

In Table 1, we report the most salient and selective neuron-concept pairs for each of our four tasks, *i.e.*, the knee points on the surprisal-selectivity Pareto fronts (Fig. 1 and 2 in Appendix C). Most

Table 1: Knee points on the surprisal-selectivity Pareto fronts. ICD code Q corresponds to congenital malformations, deformations, and chromosomal abnormalities, ICD code V to transport accidents.

Prediction target	Neuron	Layer	Concept	Top-3 features	Surprisal	Selectivity	p-value
Statistically significant results ($p < 0.05$)							
ICU transfer within 12h	113	12	Q	ICU admissions in the past year, heart rate, hypothyroidism	6.13	0.73	$2.1 \cdot 10^{-6}$
Hospitalization	49	8	V	Hospitalizations in the past three months, oxygen saturation, emergency severity index	5.80	0.62	$5.7 \cdot 10^{-3}$
Statistically insignificant results ($p \geq 0.05$)							
Inhospital mortality	78	9	V	ED visits in the past month, respiration rate, dementia	5.80	0.53	0.89
Critical outcome	62	1	V	ED visits in the past three months, pain scale, cardiac arrhythmia	7.74	0.30	1.00

knee points and indeed many points on the Pareto fronts corresponded to ICD code V (transport accidents). For instance, TabPFN neuron 49 in layer 8 may potentially encode high triage acuity (emergency severity index), hypoxemia (low oxygen saturation), and other features as signs of a traumatic chest injury due to a transport accident, which typically requires hospitalization (Table 1). We found knee points across initial, middle, and final layers in TabPFN, suggesting that concepts may not be preferentially represented at certain layers. The surprisal and selectivity values for the knee points reached statistical significance on half of our tasks ($p < 0.05$).

Overall, we provide the first evidence that interpretable neurons can emerge naturally in tabular foundation models and that they can be uncovered, at least in some cases, with a simple search across neuron-concept pairs.

5 Limitations

Although our findings provide valuable insights into the inner workings of foundation models, they should be interpreted in light of several limitations. First, our significance tests are asymptotic and assume i.i.d. mutual information and surprisal scores, which does not account for variations in empirical distributions (*e.g.*, due to data scarcity) nor for structural dependencies in the network architecture (*e.g.*, layers) or concept label space (*e.g.*, hierarchies). Second, the surprisal and selectivity scores did not reach statistical significance in all tasks, which may reflect that inpatient mortality depends on multiple interacting factors rather than single ICD codes, whereas ICU transfer or hospitalization are more strongly driven by deterministic referral policies. What is more, the empirical mutual information ranking (Eq. (1)) reduces the statistical power, so that the statistical significance is effectively determined by the selectivity (Eq. (6)). Third, the selectivity values remained at or below 0.73, meaning that we were not able to identify "grandmother cells" that (almost) exclusively respond to single concepts.

6 Conclusion

Interpreting the decision-making processes of foundation models remains an open challenge. In our work, we propose an information-theoretic testing framework that allows us to identify neurons that are both salient and selective for single concepts. Applied to the representations of TabPFN, a tabular foundation model, our framework offers the first evidence that some neurons in such models show moderate, statistically significant saliency and selectivity for high-level concepts. We are confident that our framework serves as a useful tool for mechanistic interpretability researchers and practitioners to analyze concept representations in foundation models.

Acknowledgments and Disclosure of Funding

This research was funded by the Bundesministerium für Bildung und Forschung (BMBF, project number: 16DHBKI071) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project number: 528483508). The authors declare no competing interests.

References

Adler, R., Bunte, A., Burton, S., Großmann, J., Jaschke, A., Kleen, P., Lorenz, J. M., Ma, J., Markert, K., Meeß, H., et al. (2022). Deutsche Normungsroadmap Künstliche Intelligenz.

Arora, S., Li, Y., Liang, Y., Ma, T., and Risteski, A. (2018). Linear algebraic structure of word senses, with applications to polysemy. *Transactions of the Association for Computational Linguistics*, 6:483–495.

Bertsimas, D., Pauphilet, J., and Van Parys, B. (2021). Sparse classification: a scalable discrete optimization perspective. *Machine Learning*, 110(11):3177–3209.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. (2022). On the opportunities and risks of foundation models.

Covert, I., Lundberg, S., and Lee, S.-I. (2021). Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90.

Dayan, P. and Abbott, L. F. (2005). *Theoretical neuroscience: computational and mathematical modeling of neural systems*. MIT press.

Erickson, N., Purucker, L., Tschalzev, A., Holzmüller, D., Desai, P. M., Salinas, D., and Hutter, F. (2025). TabArena: A living benchmark for machine learning on tabular data.

Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., and Lee, S.-I. (2021). Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature machine intelligence*, 3(7):620–631.

European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

Ferrando, J., Sarti, G., Bisazza, A., and Costa-jussà, M. R. (2024). A primer on the inner workings of transformer-based language models.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. (2025). Scaling and evaluating sparse autoencoders. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, Singapore. International Conference on Learning Representations (ICLR).

Gross, C. G. (2002). Genealogy of the "grandmother cell". *The Neuroscientist*, 8(5):512–518.

Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., and Bertsimas, D. (2023). Finding neurons in a haystack: Case studies with sparse probing. *Transactions on Machine Learning Research*.

Hewitt, J. and Liang, P. (2019). Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics (ACL).

Hollmann, N., Müller, S., Purucker, L., Krishnakumar, A., Körfer, M., Hoo, S. B., Schirrmeister, R. T., and Hutter, F. (2025). Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326.

Kandel, E. R., Koester, J. D., Mack, S. H., and Siegelbaum, S. A. (2021). *Principles of neural science*, volume 6. McGraw Hill.

Li, M., Jeong, S., Liu, S., and Berger, M. (2024). CAN: Concept-aligned neurons for visual comparison of deep neural network models. *Computer Graphics Forum*, 43(3):e15085.

Lieberum, T., Rajamanoharan, S., Conmy, A., Smith, L., Sonnerat, N., Varma, V., Kramar, J., Dragan, A., Shah, R., and Nanda, N. (2024). Gemma Scope: Open sparse autoencoders everywhere all at once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 278–300, Miami, USA. Association for Computational Linguistics (ACL).

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., et al. (2024). Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Nikolitsa, E. K., Kontou, P. I., and Bagos, P. G. (2025). metacp: a versatile software package for combining dependent or independent p-values. *BMC Bioinformatics*, 26(1):109.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. (2020). Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

Olver, F. W. J., Lozier, D. W., Boisvert, R. F., and Clark, C. W. (2010). *NIST handbook of mathematical functions*. Cambridge University Press.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Sharkey, L., Chughtai, B., Batson, J., Lindsey, J., Wu, J., Bushnaq, L., Goldowsky-Dill, N., Heimersheim, S., Ortega, A., Bloom, J., Biderman, S., Garriga-Alonso, A., Conmy, A., Nanda, N., Rumbelow, J., Wattenberg, M., Schoots, N., Miller, J., Michaud, E. J., Casper, S., Tegmark, M., Saunders, W., Bau, D., Todd, E., Geiger, A., Geva, M., Hoogland, J., Murfet, D., and McGrath, T. (2025). Open problems in mechanistic interpretability.

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International Conference on Machine Learning (ICML)*, pages 3145–3153. PMLR.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, pages 3319–3328. PMLR.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet. *Transformer Circuits Thread*.

Xie, F., Zhou, J., Lee, J. W., Tan, M., Li, S., Rajnthern, L. S., Chee, M. L., Chakraborty, B., Wong, A.-K. I., Dagan, A., et al. (2022). Benchmarking emergency department prediction models with machine learning and public electronic health records. *Scientific Data*, 9(1):658.

A Datasets and preprocessing

We employed the MIMIC-IV-ED dataset to tackle four tabular prediction tasks at ED triage as described by Xie et al. (2022): inhospital mortality prediction, ICU transfer within 12h prediction, critical outcome prediction (defined as either inhospital mortality or ICU transfer within 12h prediction), and hospitalization prediction, each treated as a binary classification problem. The data, collected at the Beth Israel Deaconess Medical Center between 2011 and 2019, comprised 64 features spanning patient history, demographics, vital signs, and chief complaints. The training set contained 353,150 ED episodes (samples) from 182,588 unique patients, the test set 88,287 ED episodes from 65,169 unique patients (Xie et al., 2022). We included rows with available ICD codes and extracted concepts at three hierarchical levels:

- Low-level (*e.g.*, R570 = cardiogenic shock) with 5684 concepts,
- Mid-level (*e.g.* R57 = shock) with 1239 concepts,
- High-level (*e.g.*, R = signs and symptoms) with 24 concepts.

For all tasks except hospitalization prediction, the minority class constituted $\leq 7\%$ of the training samples. To address the class imbalance, we undersampled the majority classes in the training sets for inhospital mortality, ICU transfer, and critical outcome prediction. For hospitalization prediction, where the classes were relatively balanced, we undersampled both classes to 5000 samples due to the context length limitations of our tabular foundation model (Sect. 4.1). This yielded 512, 2712, 2948, and 10000 training samples for mortality, ICU transfer, critical outcome, and hospitalization prediction, respectively.

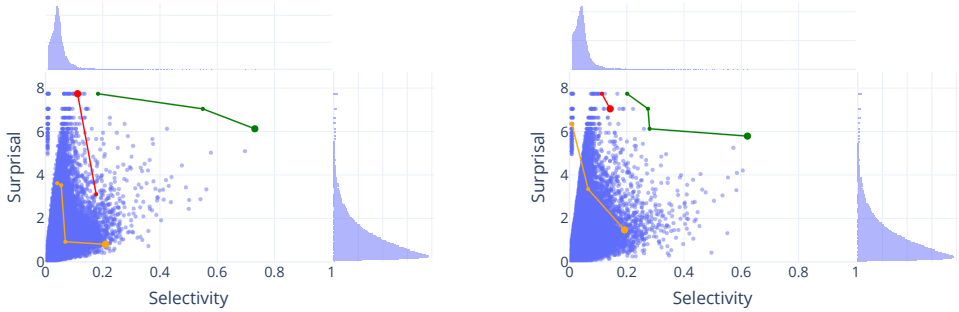
B Baselines

To contextualize the performance of our search procedure, we compared it against two sparse probing baselines. Both approaches used linear probes to predict the concept labels from the neuronal activations. The first baseline employed a local game-theoretic neuron selection method based on SHAP values (Covert et al., 2021; Lundberg and Lee, 2017), while the second used a globally optimal neuron selection strategy (Bertsimas et al., 2021; Gurnee et al., 2023):

- In the SHAP-based approach, we trained one multivariate logistic regression classifier per concept and selected the neuron with the largest mean absolute interventional SHAP value (Covert et al., 2021; Lundberg and Lee, 2017). In this case, the method is equivalent to common gradient-based explanations when mean-centering is applied, including gradient $\times$ input (Shrikumar et al., 2017), DeepLIFT (Shrikumar et al., 2017), integrated gradients (Sundararajan et al., 2017), and expected gradients (Erion et al., 2021).
- In the optimal probing approach, we trained N univariate logistic regression classifiers per concept and selected the neuron whose logistic model achieved the largest log-likelihood. This exhaustive enumeration is equivalent to using optimization solvers (Bertsimas et al., 2021; Gurnee et al., 2023), but is computationally more efficient for single-neuron selection.

For each baseline, this resulted in C selected neurons, one per concept.

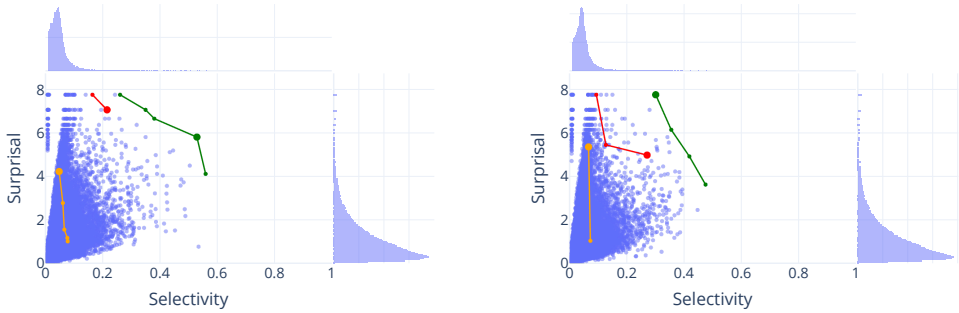
C Pareto front visualization



(a) ICU transfer within 12h prediction.

(b) Hospitalization prediction.

Figure 1: Surprisal-selectivity Pareto fronts for datasets in which the suprisal and selectivity reached statistical significance ($p < 0.05$). The Pareto fronts obtained from our search procedure, sparse probes via SHAP values (Covert et al., 2021; Lundberg and Lee, 2017), and optimal probing (Bertsimas et al., 2021; Gurnee et al., 2023) are shown in green, orange, and red, respectively. Larger markers indicate knee points. Both sparse probes via SHAP values and optimal probing are Pareto-dominated by our method. Values < 0.01 are omitted for improved readibility.



(a) Inhospital mortality prediction.

(b) Critical outcome prediction.

Figure 2: Surprisal-selectivity Pareto fronts for datasets in which the suprisal and selectivity did not reach statistical significance ($p \geq 0.05$). The Pareto fronts obtained from our search procedure, sparse probes via SHAP values (Covert et al., 2021; Lundberg and Lee, 2017), and optimal probing (Bertsimas et al., 2021; Gurnee et al., 2023) are shown in green, orange, and red, respectively. Larger markers indicate knee points. Both sparse probes via SHAP values and optimal probing are Pareto-dominated by our method. Values < 0.01 are omitted for improved readibility.