ConceptTracer: Interactive Analysis of Concept Saliency and Selectivity in Neural Representations

Ricardo Knauer[1,*], Andre Beinrucker[1] and Erik Rodner[1]

[1]KI-Werkstatt, University of Applied Sciences Berlin, Berlin, Germany

Abstract

Neural networks deliver impressive predictive performance across a variety of tasks, but they are often opaque in their decision-making processes. Despite a growing interest in mechanistic interpretability, tools for systematically exploring the representations learned by neural networks in general, and tabular foundation models in particular, remain limited. In this work, we introduce ConceptTracer, an interactive application for analyzing neural representations through the lens of human-interpretable concepts. ConceptTracer integrates two information-theoretic measures that quantify concept saliency and selectivity, enabling researchers and practitioners to identify neurons that respond strongly to individual concepts. We demonstrate the utility of ConceptTracer on representations learned by TabPFN and show that our approach facilitates the discovery of interpretable neurons. Together, these capabilities provide a practical framework for investigating how neural networks like TabPFN encode concept-level information. ConceptTracer is available at https://github.com/ml-lab-htw/concept-tracer.

Keywords

mechanistic interpretability, concept-based explainability, representation analysis

1. Introduction

Neural networks achieve remarkable predictive performance across a wide range of tasks, yet the mechanisms underlying their predictions are often difficult to interpret [1, 2]. This opacity can hinder their deployment in domains where transparency is critical [3, 4]. Developing a deeper understanding of their learned representations can improve model interpretability and strengthen user trust, particularly in high-stakes environments such as clinical practice [3, 4]. While mechanistic interpretability research has introduced advanced methods for revealing the inner workings of neural networks, *e.g.*, by identifying interpretable units of computation [5, 6], tools for systematically analyzing the representations of neural networks in general and tabular foundation models in particular remain scarce. As a result, researchers and practitioners often lack practical frameworks for examining how neural networks encode information and whether these representations align with higher-order, human-understandable features, *i.e.*, concepts. Analyzing how such concepts are encoded within a model provides a principled way to investigate what information internal representations capture and to what extent they reflect human-interpretable structure.

Our contributions are as follows:

- We introduce ConceptTracer, an interactive system for analyzing neural representations grounded in human-understandable concepts (Sect. 4). ConceptTracer incorporates two information-theoretic measures that quantify the neural saliency and selectivity for individual concepts [7, 8, 9] (Sect. 3).
- We demonstrate the practical value of ConceptTracer on representations learned by TabPFN [10, 11]. Our results show that our approach supports the identification of interpretable neurons (Sect. 5).

Late-breaking work, Demos and Doctoral Consortium, colocated with the 4th World Conference on eXplainable Artificial Intelligence: July 01–03, 2026, Fortaleza, Brazil

*Corresponding author.

ricardo.knauer@htw-berlin.de (R. Knauer); andre.beinrucker@htw-berlin.de (A. Beinrucker); erik.rodner@htw-berlin.de (E. Rodner)

0009-0003-9258-5105 (R. Knauer)

2. Related work

A central question in mechanistic interpretability is whether individual neurons are directly interpretable because they respond strongly to single concepts, an idea linked to the "grandmother cell" hypothesis in neuroscience [12, 13]. However, frameworks for analyzing neural representations at the level of individual neurons and concepts remain underexplored relative to established explainability tools [14, 15] and recent methods that localize and disentangle distributed and superposed representations via auxiliary model training [16, 17, 18, 19, 20]. To the best of our knowledge, Neuronpedia [21] is currently the only application that supports an interactive exploration and visualization of neural representations, even though its functionality centers on features learned by sparse autoencoders and concepts identified through unsupervised discovery, which require substantial computational resources, rather than neurons in the underlying model and predefined concepts.

In the next section, we introduce two metrics to analyze neural representations at the level of individual model neurons and ground-truth concepts [7, 8, 9] using information theory [22, 23, 24].

3. Quantifying neural interpretability

In this section, we introduce our two information-theoretic measures for neural interpretability. Let $\mathbf{A} \in \mathbb{R}^{M \times N}$ be the neural activations for $M \in \mathbb{N}_+$ samples and $N \in \mathbb{N}_+$ neurons. The network structure is irrelevant for the following definitions and is thus omitted. Moreover, let $\mathbf{B} \in \{0,1\}^{M \times C}$ represent the concept labels for the M samples and $C \in \mathbb{N}_+$ higher-order, human-interpretable features, *i.e.*, concepts. To quantify 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, we employ the normalized mutual information $\hat{I}(\mathbf{a}_i, \mathbf{b}_j) \in [0,1]$ as a nonlinear measure of statistical dependence [22, 23, 24]. We define the saliency [8, 9] for neuron i with respect to concept j as this mutual information:

$$\text{saliency}(\mathbf{a}_i, \mathbf{b}_j) = \hat{I}(\mathbf{a}_i, \mathbf{b}_j), \quad 0 \leq \text{saliency}(\mathbf{a}_i, \mathbf{b}_j) \leq 1 \quad . \tag{1}$$

The saliency thus captures the strength of association between a neuron and a concept. However, it does not indicate whether a neuron is specialized for a particular concept. To capture this specialization, we define the selectivity [7, 8] as the fraction of a neuron's total saliency that is attributable to a given concept:

$$\text{selectivity}(\mathbf{a}_i, \mathbf{b}_j) = \frac{\text{saliency}(\mathbf{a}_i, \mathbf{b}_j)}{\sum_{c=1}^{C} \text{saliency}(\mathbf{a}_i, \mathbf{b}_c)}, \quad 0 \leq \text{selectivity}(\mathbf{a}_i, \mathbf{b}_j) \leq 1 \quad . \tag{2}$$

By construction, the selectivity depends on the size of the concept set: smaller concept sets tend to yield higher selectivity values due to the normalization. Moreover, the selectivity only accounts for the concepts $c \in \{1, ..., C\}$ under consideration. To determine whether our information-theoretic measures exceed what would be expected by chance, we employ nonparametric permutation testing [25] to estimate empirical null distributions for both metrics. Our null hypothesis H_0 is that the neural activations and concept labels are independent, *i.e.*, no neuron is preferentially associated with any concept and no concept is preferentially associated with any neuron. Permutation tests are well suited in this setting because they only require that samples are exchangeable under permutations of the concept labels, while parametric tests typically rely on null distributions that are derived under much stronger assumptions [26, 27]. We generate permutations by applying a single shuffle of the sample indices across all concept labels, thereby preserving structural dependencies within the concept label space, and count the number of permutations in which the permuted test statistic equals or exceeds the observed value. This procedure yields empirical p-values for the saliency and the selectivity, $p_{saliency}(\mathbf{a}_i, \mathbf{b}_j)$ and $p_{selectivity}(\mathbf{a}_i, \mathbf{b}_j)$. To control the family-wise error rate, we apply a max-statistic correction [28] to both p-values and subsequently combine them into $p_{combined}(\mathbf{a}_i, \mathbf{b}_j)$ using a Bonferroni correction [29].

In the next section, we show how concept saliency and selectivity can be integrated into an interactive mechanistic interpretability dashboard to facilitate the discovery of interpretable neurons in neural representations.

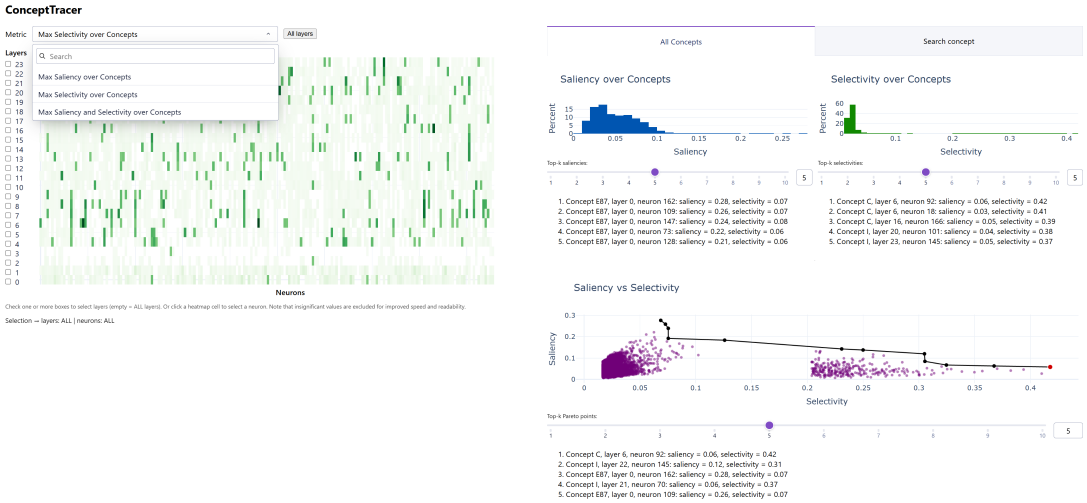


Figure 1: The ConceptTracer dashboard for the interactive analysis of neural representations.

4. ConceptTracer

In this section, we introduce ConceptTracer[1], an interactive application for analyzing neural representations through the lens of human-interpretable concepts that is built around four main components:

- *Dataset*: A set of training and test samples.
- *Neural network*: A model for training and testing.
- *Embedding extractor*: A function for extracting neural representations from the trained model.
- *Concepts*: A set of higher-order, human-interpretable features.

Once the components are specified, ConceptTracer computes the concept saliency and selectivity, along with their individual and combined p-values, for any neuron-concept pair (Sect. 3). These measures can then be systematically explored through the ConceptTracer dashboard (Fig. 1). By default, only significant values at a significance level of 0.05 are displayed.

The dashboard consists of two main panels. The left panel enables users to explore the maximum saliency, selectivity, or combined score across concepts at different levels of the network: network-wide, per layer, or per neuron. The right panel visualizes the distribution of the metrics within the chosen subnetwork and highlights the top-k neuron-concept pairs. To facilitate the identification of pairs that are both salient and selective, the Pareto front is displayed. In addition, the knee point on the Pareto front is marked to indicate the most salient and selective pair. The knee point is determined by maximizing the sum of the min-max scaled saliency and selectivity scores. Beyond concept-wide exploration, users can also search for specific concepts and perform analyses at the level of individual concepts through the `Search concept` tab. ConceptTracer therefore enables the examination of neuron-concept associations from four complementary perspectives:

- *Network view*: How salient and selective are the neurons across the entire network for the set of concepts?
- *Layer view*: How salient and selective are the neurons within a selected layer or set of layers for the set of concepts?
- *Neuron view*: How salient and selective is a selected neuron for the set of concepts?
- *Concept view*: How salient and selective are the neurons in the network, selected layer, or set of layers for a particular concept?

This allows for a comprehensive investigation of how neural networks encode concept-level information. In the next section, we empirically test whether we can use ConceptTracer to find interpretable neurons in neural representations.

[1]https://github.com/ml-lab-htw/concept-tracer

5. Experiments

In this section, we use our definitions of concept saliency and selectivity (Sect. 3) within ConceptTracer (Sect. 4) to evaluate whether our framework can uncover interpretable neurons in neural representations. To this end, we apply ConceptTracer on representations learned by TabPFN [10, 11] and demonstrate that our approach supports the identification of interpretable neurons.

5.1. Experimental setup

5.1.1. Dataset and tasks

Understanding model behavior is particularly important for mitigating risks in high-stakes domains such as clinical practice [3, 4]. Therefore, we conducted our experiments using the MIMIC-IV-ED dataset to solve four emergency department (ED) triage tasks: inhospital mortality prediction, intensive care unit (ICU) transfer within 12h prediction, critical outcome prediction (defined as either inhospital mortality or ICU transfer within 12h), and hospitalization prediction [30]. Each task was treated as a binary classification problem. The data was collected at the Beth Israel Deaconess Medical Center between 2011 and 2019 and encompassed 64 features, spanning patient history, demographics, vital signs, chief complaints, and comorbidities. 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 [30]. Concepts were defined using diagnostic codes from the International Classification of Diseases (ICD) organized at three hierarchical levels:

- *High-level*: *e.g.*, R = signs and symptoms.
- *Mid-level*: *e.g.*, R57 = shock.
- *Low-level*: *e.g.*, R570 = cardiogenic shock.

5.1.2. Data and concept preprocessing

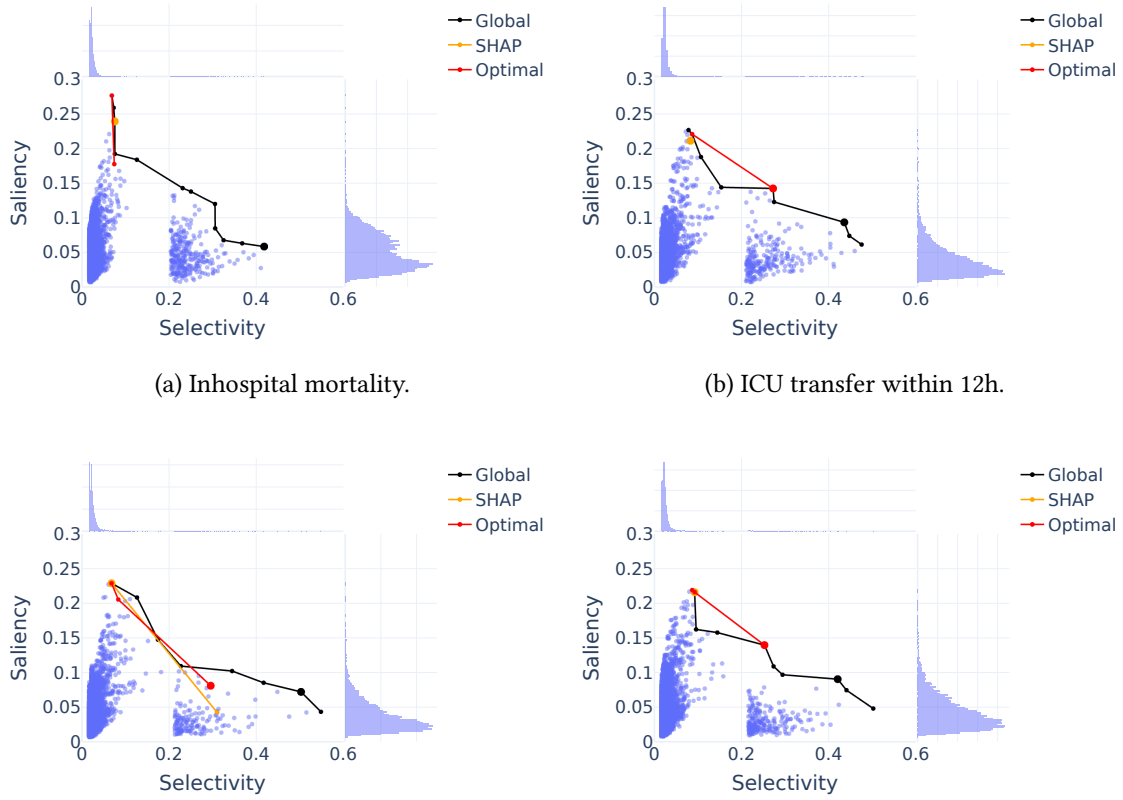
We excluded training set patients who also appeared in the test set to prevent data leakage and removed rows without available ICD codes. For all tasks except hospitalization prediction, the minority class constituted $\leq 8\%$ of the training samples. To address the class imbalance, we undersampled the majority classes in the training sets, resulting in 352, 1,318, 1,498, and 7,512 training samples for mortality, ICU transfer, critical outcome, and hospitalization prediction, respectively. For the low- and mid-level concepts, the concept prevalence followed a long-tailed distribution, with a large fraction of concepts supported by very few or even single samples. To mitigate the concept imbalance, we filtered out concepts with a prevalence < 100, yielding 354, 282, and 24 low-, mid-, and high-level concepts, respectively.

5.1.3. Model and evaluation

We employed TabPFN 6.3.2 [10, 11] as our neural network due to its strong reported performance in tabular prediction tasks [31]. TabPFN was used without tuning or ensembling and its discriminative performance was evaluated using the test set area under the receiver operating characteristic curve (AUC). Next, we extracted the 192-dimensional target-column embeddings [11] from each of its 24 encoder layers [10] and passed them, together with the concept labels, to ConceptTracer to systematically analyze the neuron-concept pairs. As baselines, we employed sparse probes based on SHAP values [32, 33] and optimal probing [16, 18]. Please refer to Knauer and Rodner [26] for implementation details on the baselines.

5.2. Experimental results

TabPFN achieved test set AUC values ranging from 0.79 to 0.85. Less than 1% of the neurons were significantly salient and selective for specific concepts. Due to the normalization in the selectivity



(a) Inhospital mortality.

(b) ICU transfer within 12h.

(c) Critical outcome.

(d) Hospitalization.

Figure 2: Significant neuron-concept pairs for our tasks. Black denotes the global Pareto front, with larger markers indicating knee points. The Pareto fronts for the sparse probing baselines via SHAP values and optimal probing are shown in orange and red, respectively.

calculation (Eq. 2), the significant neuron-concept pairs formed two distinct clusters of low- and mid-level concepts with smaller selectivity values versus high-level concepts with larger selectivity values (Fig. 2). Sparse probes via SHAP values and optimal probing only partially recovered the global Pareto fronts and missed the most selective pairs (Fig. 2). This suggests that identifying the most salient and selective neuron-concept pairs requires an exhaustive search rather than relying solely on sparse probing approaches.

The global Pareto fronts included neuron-concept pairs from initial, middle, and final layers of TabPFN, indicating that concepts may not be preferentially represented in particular layers. The neurons on the global Pareto fronts were associated with the following concepts:

- *Inhospital mortality*: C (malignant neoplasms), E66 (obesity), E87 (other disorders of fluid, electrolyte, and acid-base balance), I (diseases of the circulatory system).
- *ICU transfer within 12h*: F (mental and behavioral disorders), F32 (depressive episode).
- *Critical outcome*: F, F32.
- *Hospitalization*: F, F32, I50 (heart failure).

These results suggest two complementary neural response patterns that are consistent with concept differentiation within composite input features and concept integration across related input features. For example, neuron 162 in layer 0 on the Pareto front shows salient and selective responses to ICD code E87 (other disorders of fluid, electrolyte, and acid-base balance). In contrast, its saliency and selectivity for other subconcepts within the fluid and electrolyte disorders feature are substantially reduced, *e.g.*, by 0.12 and 0.03 for ICD code E86 (volume depletion). This pattern is consistent with differentiation

within a composite input feature. Conversely, neuron 145 in layer 22 on the Pareto front shows salient and selective responses to ICD code I (diseases of the circulatory system). When conditioned on the presence of ICD code I, cardiac arrythmias and hypertension rank among the top-4 features by mutual information with the neuron's activations. This pattern is consistent with integration across related input features.

Overall, our results demonstrate that ConceptTracer facilitates the discovery of interpretable neurons and provides a framework for systematically exploring how neural networks encode concept-level information.

6. Conclusion

Interpreting the decision-making processes of neural networks remains a big challenge. To address this challenge, we introduce ConceptTracer, an interactive application for analyzing neural representations by relating them to human-interpretable concepts. We show that ConceptTracer supports the identification of interpretable neurons in the tabular foundation model TabPFN and facilitates deeper exploration of how the network encodes concept-level information.

Several promising directions emerge from our work. For a more fine-grained analysis, future work could extend ConceptTracer by decomposing the association between neural representations and concepts into unique, redundant, and synergistic components via partial information decomposition [34]. For bias detection and mitigation, our framework could be applied to find bias-related neurons and reduce their influence via activation steering [35]. Finally, ConceptTracer could help to better understand and deliberately shape the inductive biases of tabular foundation models toward meaningful real-world concepts [10]. We believe that our framework provides a valuable resource for practitioners and researchers working in mechanistic interpretability.

Acknowledgments

This research was funded by the European Union's Erasmus+ program (project: EUonAIR, project number: 101177370) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project: Berlin Initiative for Applied Foundation Model Research, project number: 528483508).

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5 for dashboard prototyping. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the content.

References

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

K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, P. Liang, On the opportunities and risks of foundation models, 2022. URL: https://arxiv.org/abs/2108.07258. arXiv:2108.07258.

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

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

[4] European Parliament and Council of the European Union, 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), 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[5] J. Ferrando, G. Sarti, A. Bisazza, M. R. Costa-jussà, A primer on the inner workings of transformer-based language models, 2024. URL: https://arxiv.org/abs/2405.00208. arXiv:2405.00208.

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

[7] J. Hewitt, P. Liang, 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), Association for Computational Linguistics (ACL), Hong Kong, China, 2019, pp. 2733–2743. URL: https://aclanthology.org/D19-1275/. doi:10.18653/v1/D19-1275.

[8] E. R. Kandel, J. D. Koester, S. H. Mack, S. A. Siegelbaum, Principles of neural science, volume 6, McGraw Hill, 2021.

[9] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014. URL: https://arxiv.org/abs/1312.6034. arXiv:1312.6034.

[10] L. Grinsztajn, K. Flöge, O. Key, F. Birkel, P. Jund, B. Roof, B. Jäger, D. Safaric, S. Alessi, A. Hayler, M. Manium, R. Yu, F. Jablonski, S. B. Hoo, A. Garg, J. Robertson, M. Bühler, V. Moroshan, L. Purucker, C. Cornu, L. C. Wehrhahn, A. Bonetto, B. Schölkopf, S. Gambhir, N. Hollmann, F. Hutter, TabPFN-2.5: Advancing the state of the art in tabular foundation models, 2026. URL: https://arxiv.org/abs/2511.08667. arXiv:2511.08667.

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

[12] C. G. Gross, Genealogy of the "grandmother cell", The Neuroscientist 8 (2002) 512–518.

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

[14] O. Dijk, oegesam, R. Bell, Lily, Simon-Free, B. Serna, E. Ferdman, rajgupt, yanhong-zhao-ef, A. Gädke, A. Todor, A. Kulkarni, Evgeniy, Hugo, J. Salomon, M. Haizad, S. Soni, T. Okumus, woochan-jang, explainerdashboard, 2026. URL: https://doi.org/10.5281/zenodo.18526511.

[15] Microsoft, Responsible AI dashboard, 2026. URL: https://learn.microsoft.com/en-us/azure/machine-learning/concept-responsible-ai-dashboard.

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

[17] L. Gao, T. D. la Tour, H. Tillman, G. Goh, R. Troll, A. Radford, I. Sutskever, J. Leike, J. Wu, Scaling and evaluating sparse autoencoders, in: Proceedings of the 13th International Conference on Learning Representations (ICLR), International Conference on Learning Representations (ICLR), Singapore, 2025. URL: https://openreview.net/forum?id=tcsZt9ZNKD.

[18] W. Gurnee, N. Nanda, M. Pauly, K. Harvey, D. Troitskii, D. Bertsimas, Finding neurons in a haystack: Case studies with sparse probing, Transactions on Machine Learning Research (2023).
[19] T. Lieberum, S. Rajamanoharan, A. Conmy, L. Smith, N. Sonnerat, V. Varma, J. Kramar, A. Dragan, R. Shah, N. Nanda, 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, Association for Computational Linguistics (ACL), Miami, USA, 2024, pp. 278–300. URL: https://aclanthology.org/2024.blackboxnlp-1.19/. doi:10.18653/v1/2024.blackboxnlp-1.19.
[20] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, T. Henighan, Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet, Transformer Circuits Thread (2024). URL: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.
[21] J. Lin, Neuronpedia: Interactive reference and tooling for analyzing neural networks, 2023. URL: https://www.neuronpedia.org.
[22] P. Dayan, L. F. Abbott, Theoretical neuroscience: computational and mathematical modeling of neural systems, MIT press, 2005.
[23] T. Oikarinen, T.-W. Weng, CLIP-dissect: Automatic description of neuron representations in deep vision networks, in: Proceedings of the 11th International Conference on Learning Representations (ICLR), International Conference on Learning Representations (ICLR), Kigali, Rwanda, 2023. URL: https://openreview.net/forum?id=iPWiwWHc1V.
[24] C. E. Shannon, A mathematical theory of communication, The Bell System Technical Journal 27 (1948) 379–423.
[25] D. François, V. Wertz, M. Verleysen, The permutation test for feature selection by mutual information., in: Proceedings of the 14th European Symposium on Artificial Neural Networks (ESANN), European Symposium on Artificial Neural Networks (ESANN), Bruges, Belgium, 2006, pp. 239–244. URL: https://www.esann.org/proceedings/2006.
[26] R. Knauer, E. Rodner, In search of grandmother cells: Tracing interpretable neurons in tabular representations, 2026. URL: https://arxiv.org/abs/2601.03657. arXiv:2601.03657.
[27] R. A. Ince, B. L. Giordano, C. Kayser, G. A. Rousselet, J. Gross, P. G. Schyns, A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula, Human brain mapping 38 (2017) 1541–1573.
[28] P. H. Westfall, S. S. Young, Resampling-based multiple testing: Examples and methods for p-value adjustment, John Wiley & Sons, 1993.
[29] E. K. Nikolitsa, P. I. Kontou, P. G. Bagos, metacp: a versatile software package for combining dependent or independent p-values, BMC Bioinformatics 26 (2025) 109.
[30] F. Xie, J. Zhou, J. W. Lee, M. Tan, S. Li, L. S. Rajnthern, M. L. Chee, B. Chakraborty, A.-K. I. Wong, A. Dagan, et al., Benchmarking emergency department prediction models with machine learning and public electronic health records, Scientific Data 9 (2022) 658.
[31] N. Erickson, L. Purucker, A. Tschalzev, D. Holzmüller, P. M. Desai, D. Salinas, F. Hutter, TabArena: A living benchmark for machine learning on tabular data, Advances in neural information processing systems 39 (2025). URL: https://openreview.net/forum?id=jZqCqpCLdU.
[32] I. Covert, S. Lundberg, S.-I. Lee, Explaining by removing: A unified framework for model explanation, Journal of Machine Learning Research 22 (2021) 1–90.
[33] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 31 (2017). URL: https://dl.acm.org/doi/10.5555/3295222.3295230.
[34] P. L. Williams, R. D. Beer, Nonnegative decomposition of multivariate information, 2010. URL: https://arxiv.org/abs/1004.2515. arXiv:1004.2515.
[35] S. Dev, T. Li, J. M. Phillips, V. Srikumar, On measuring and mitigating biased inferences of word embeddings, in: Proceedings of the 44th AAAI conference on artificial intelligence, AAAI Press, New York, USA, 2020, pp. 7659–7666. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6267/6123.