Interpreting Latent Representations of Histopathology Foundation Models

Ing. Šimon Šmída

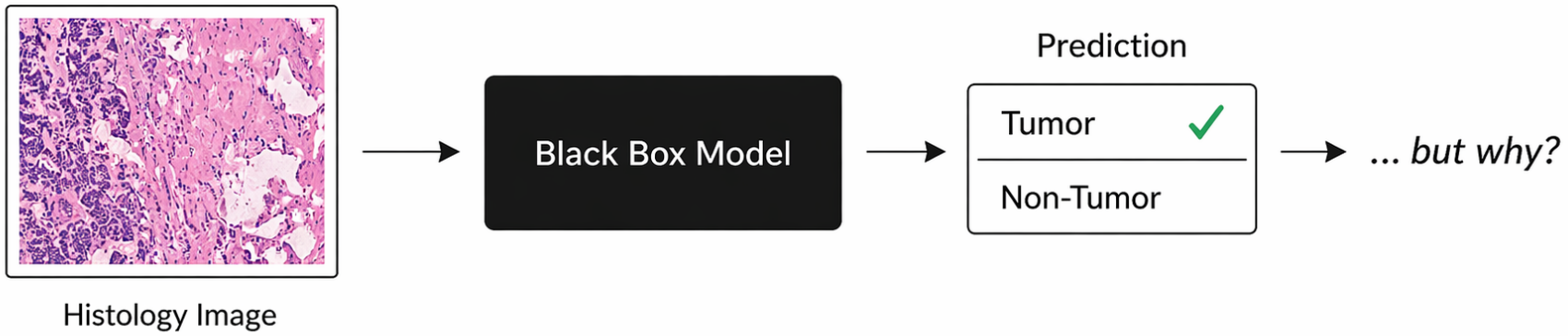
Supervisor: prof. Ing. Vanda Benešová, CSc.



FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Motivation: sometimes we need *more* than just predictions..

- Deep learning models are used in **high-stakes domains**
  - Medical AI (e.g., histology) for tumor classification, detection, …

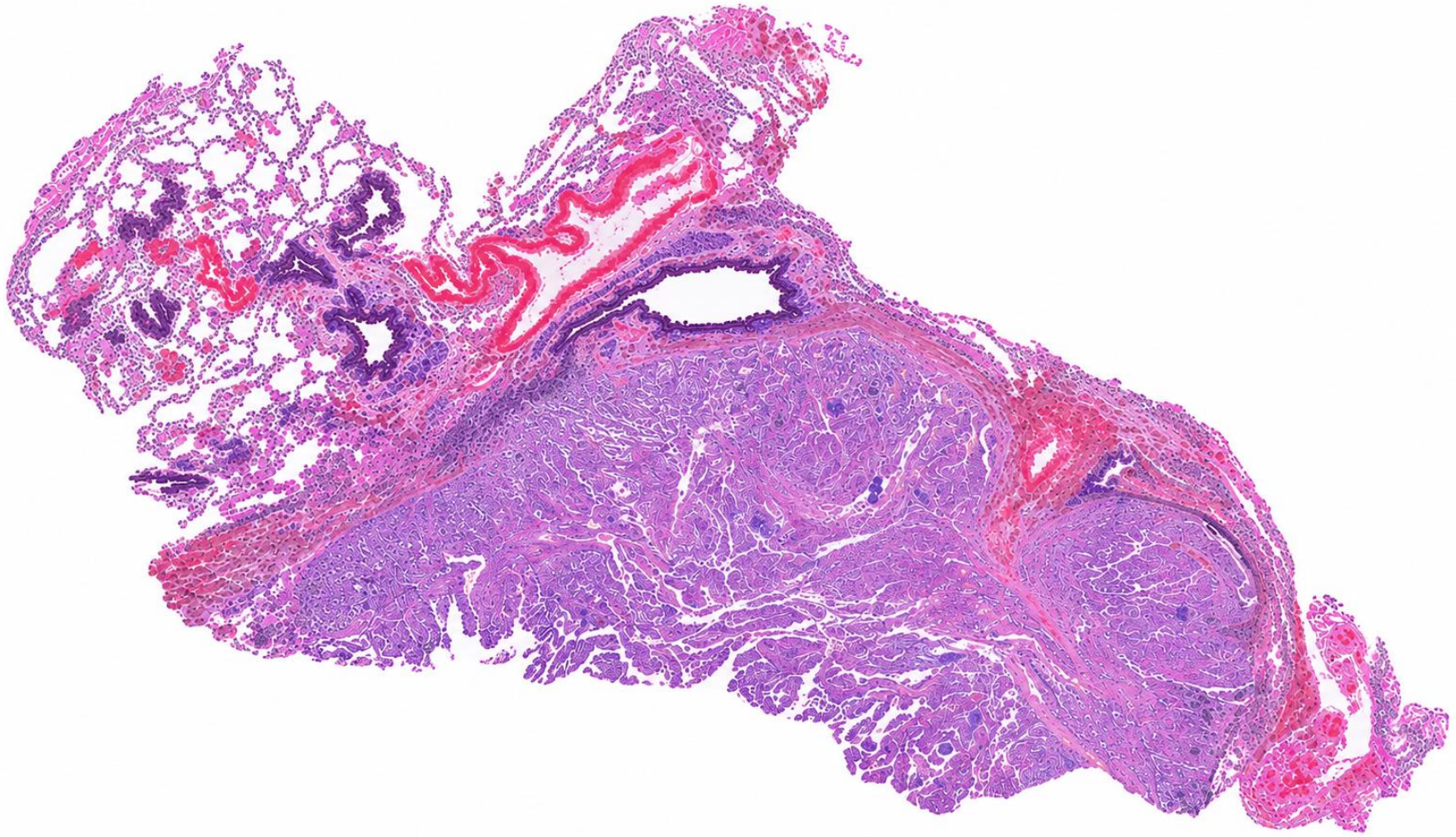


- But deep neural networks (DNNs) are considered to be *"black boxes"*
  - Predictions alone are *not enough*

- *How* DNNs make their predictions? *What structure* they learn?
- *When* can they *fail* in practice*? Why?*

Motivation: The Black Box in Computational Pathology

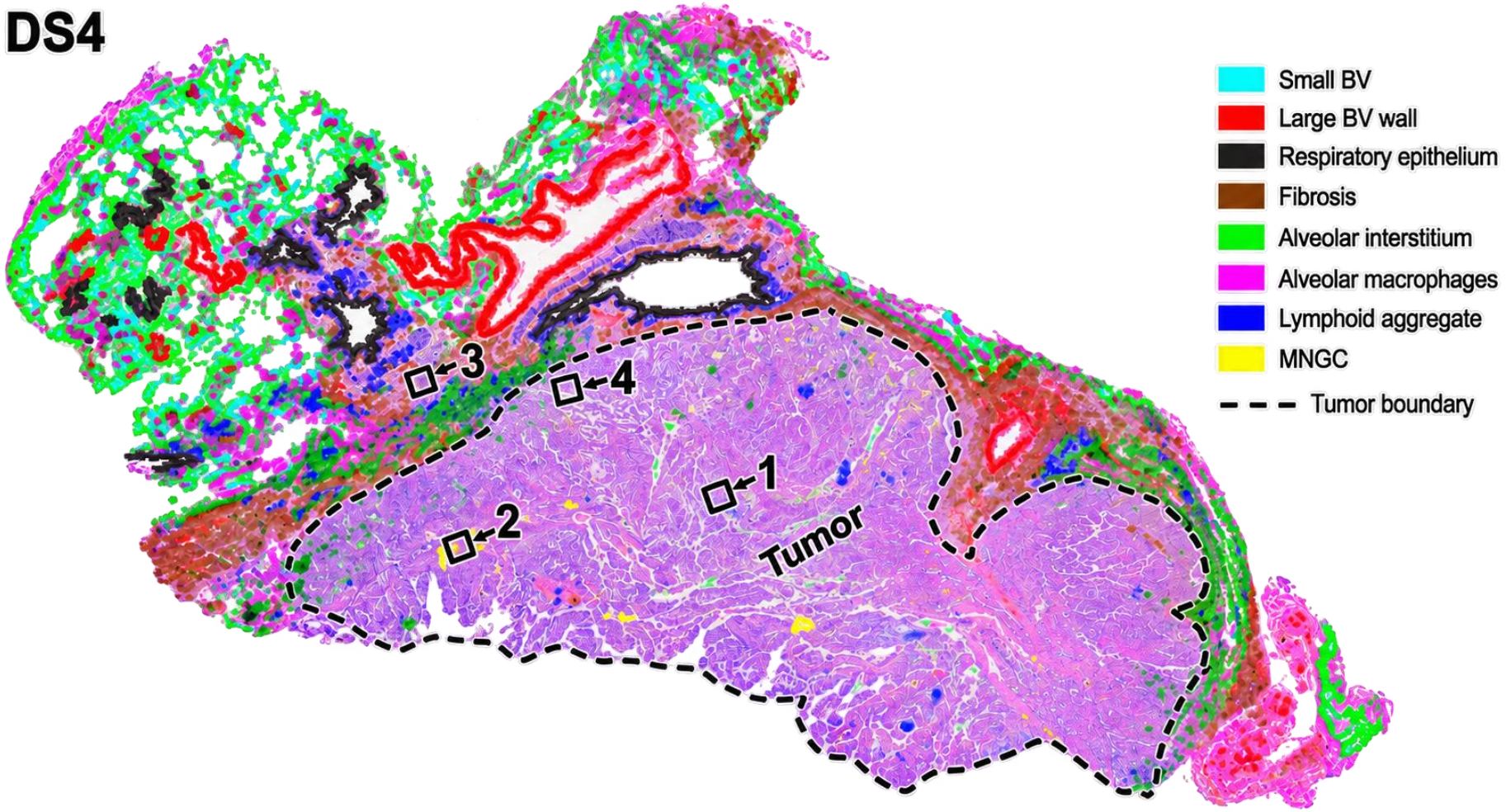
- Can we understand what AI “sees” in the tissue?



Altered from: Zhao et al. (2026). *Compositional and interpretable representation of histology using AI foundation models and sparse autoencoders. bioRxiv.* https://doi.org/10.64898/2026.06.03.725182

Motivation: The Black Box in Computational Pathology

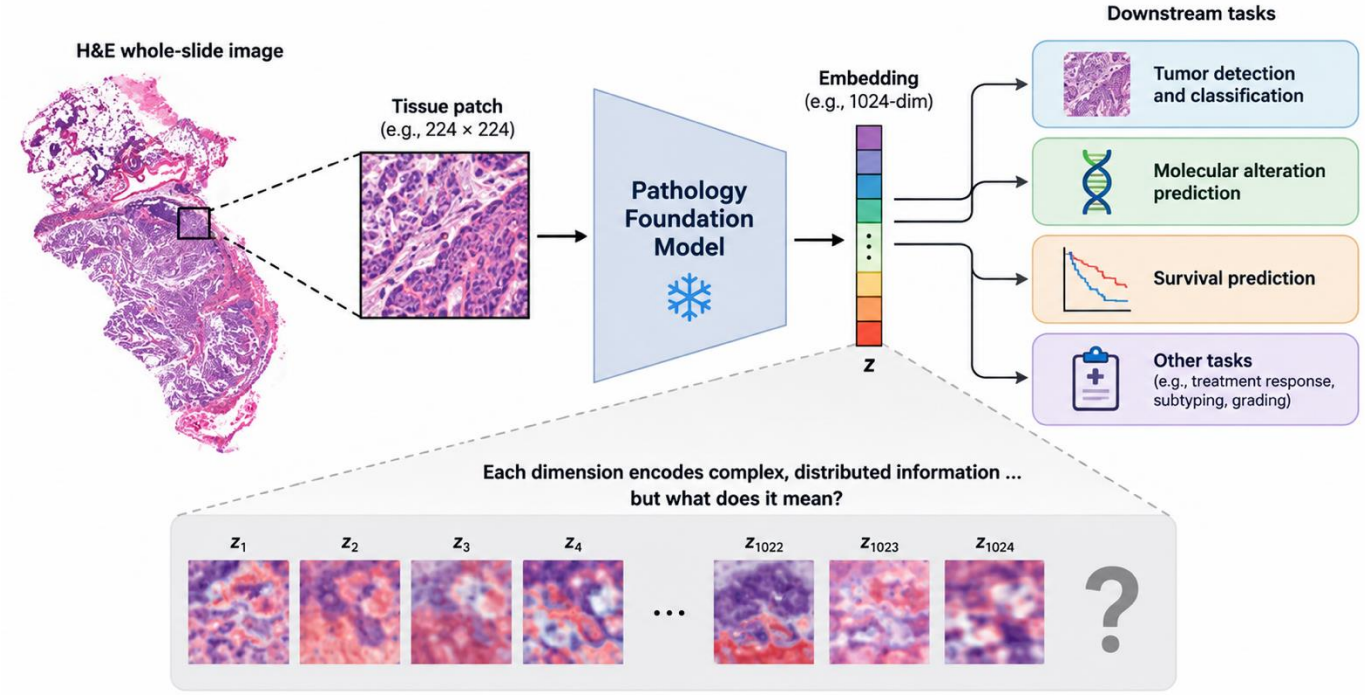
- Can we understand what AI “sees” in the tissue?



Altered from: Zhao et al. (2026). *Compositional and interpretable representation of histology using AI foundation models and sparse autoencoders. bioRxiv.* https://doi.org/10.64898/2026.06.03.725182

Foundation Models: powerful representations, unclear meaning

- Pathology Foundation Models (PFM): CONCH[1], UNI2-h[2]
- Patch -> PFM -> embedding



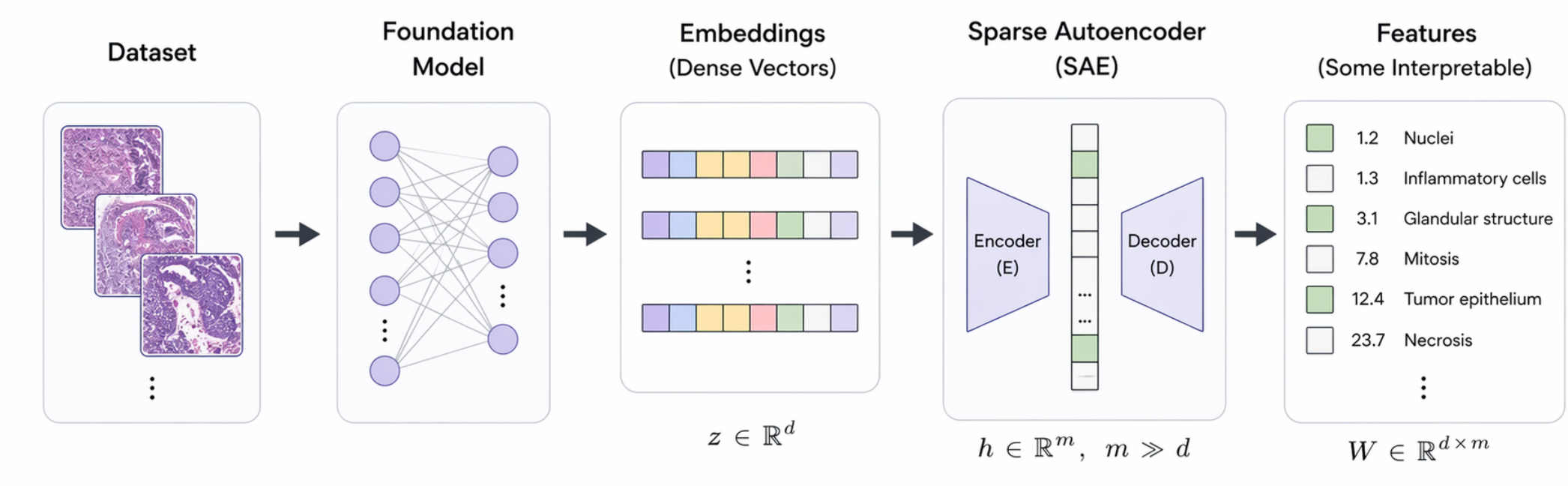
- Are *raw* embeddings interpretable? we observe *polysemanticity*

[1] Lu et al. (2024). *A visual-language foundation model for computational pathology. Nature Medicine.*

[2] Chen et al. (2024). *Towards a general-purpose foundation model for computational pathology. Nature Medicine. (Original UNI paper; UNI2-h is a later release.)*

Can we discover meaningful concepts inside representations?

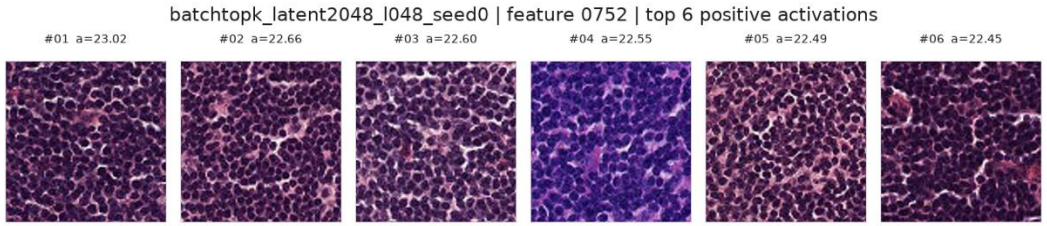
- Observation: polysemanticity
  - Proposed explanation: superposition hypothesis (assumption for SAEs)
- Can we learn new, more interpretable representation space?
  - Sparse Autoencoders (SAEs)
  - Decomposing embeddings into a small number of reusable features
- Dense embeddings vs. sparse embeddings



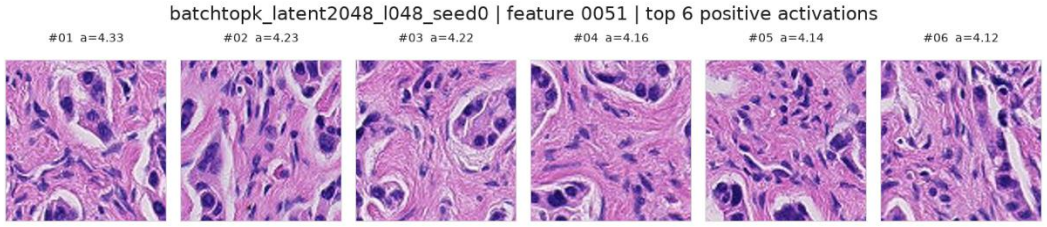
My Research: SAEs for expert-assisted analysis of PFM

- Select a feature of interest and visualize **top-activating patches**
- SAE features show visually coherent histological patterns (6 examples below)

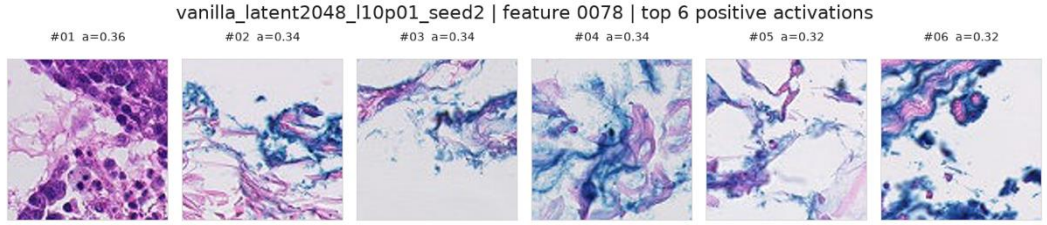
1.



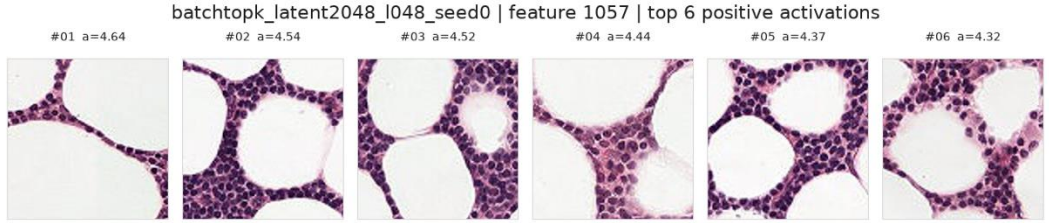
2.



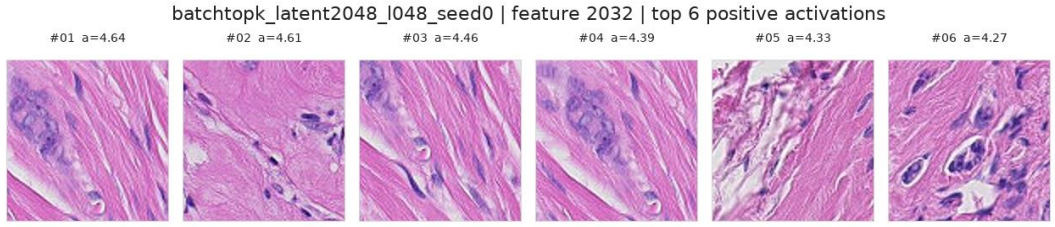
3.



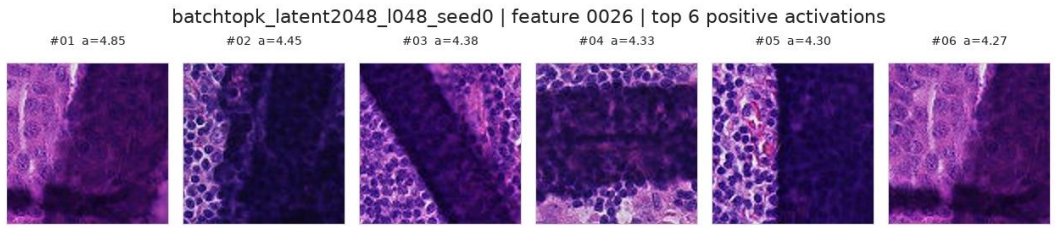
4.



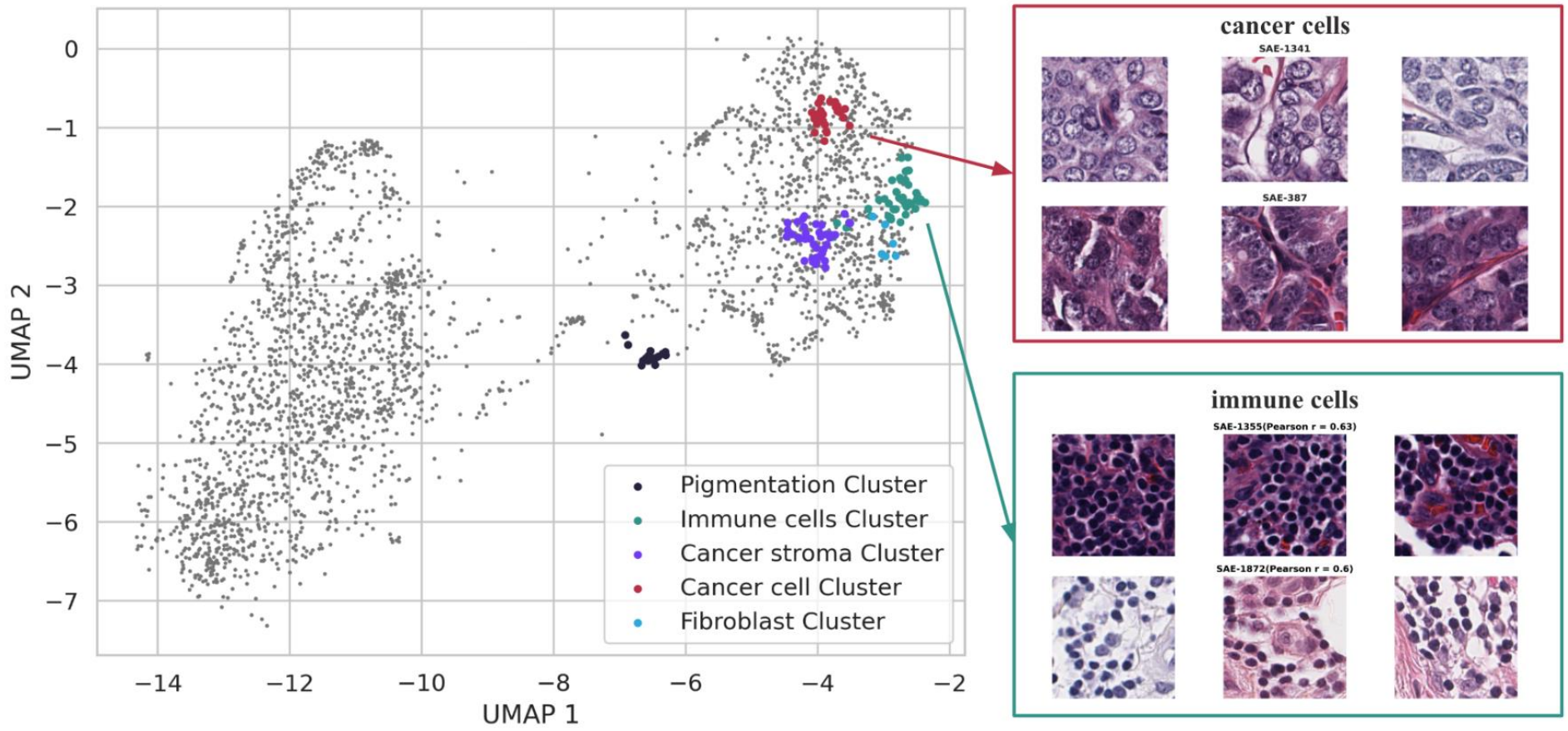
5.



6.

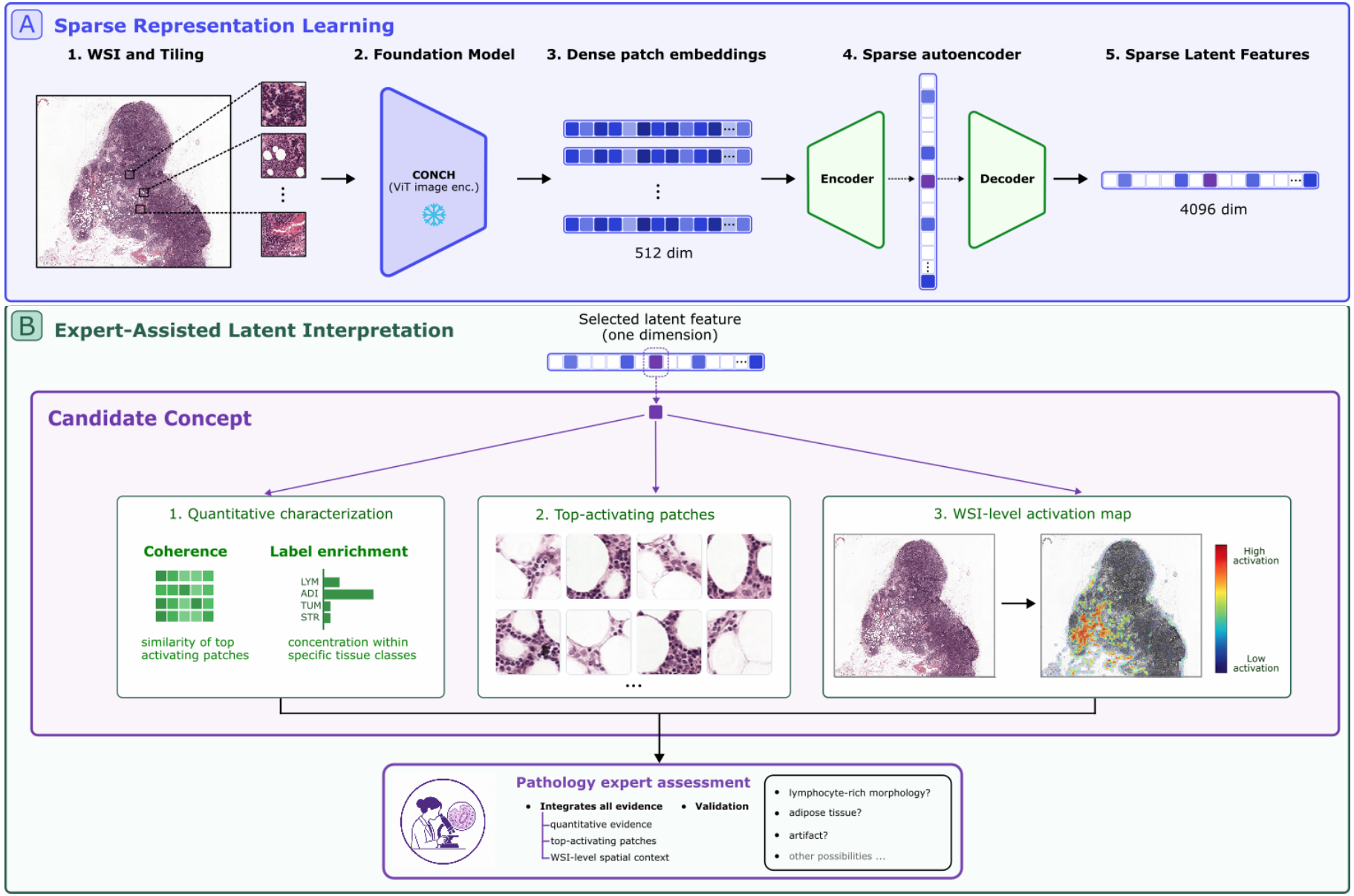


- Not simple clustering



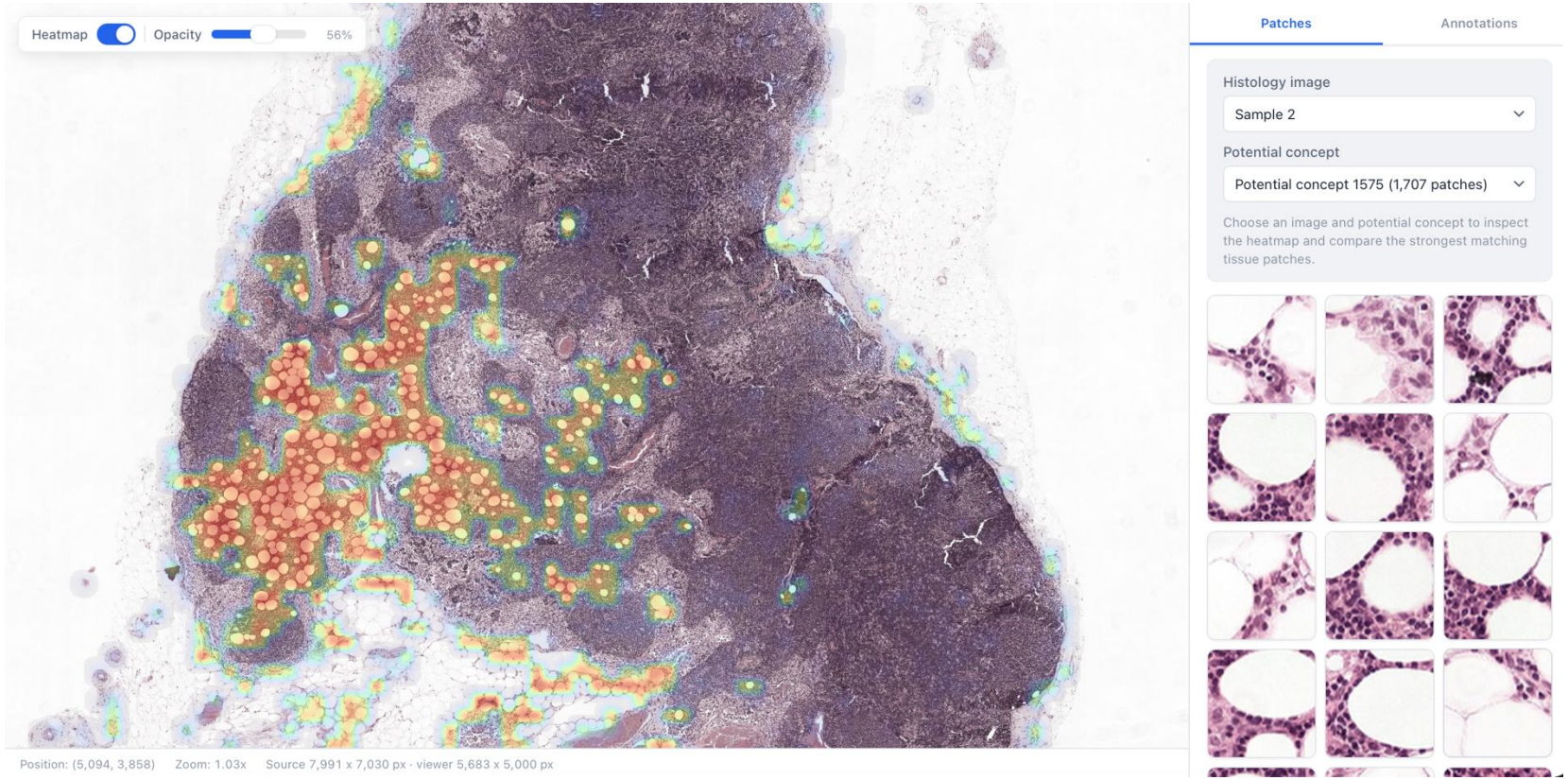
Source: Le et al. (2024). *Learning biologically relevant features in a pathology foundation model using sparse autoencoders. NeurIPS 2024 Workshop.*

My Research: SAEs for expert-assisted analysis of PFM



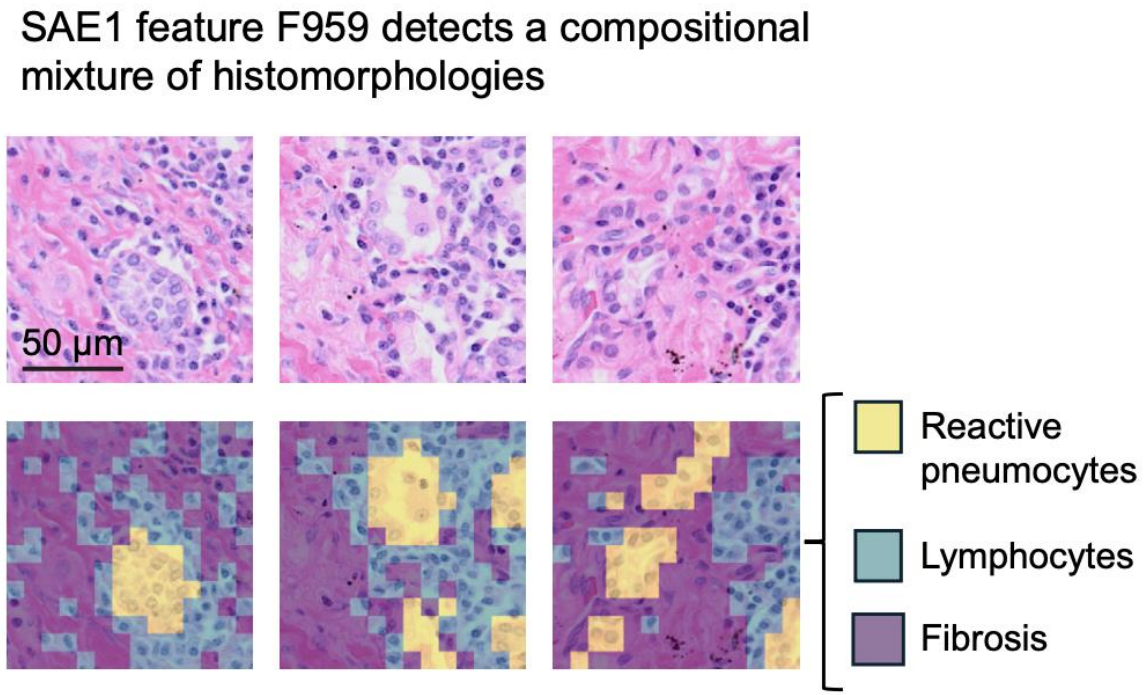
Spatial Maps of Top Activations per Feature

- Spatial activations on the whole-slide tissue image (WSI)
- Map features back to the tissue to see where they activate spatially



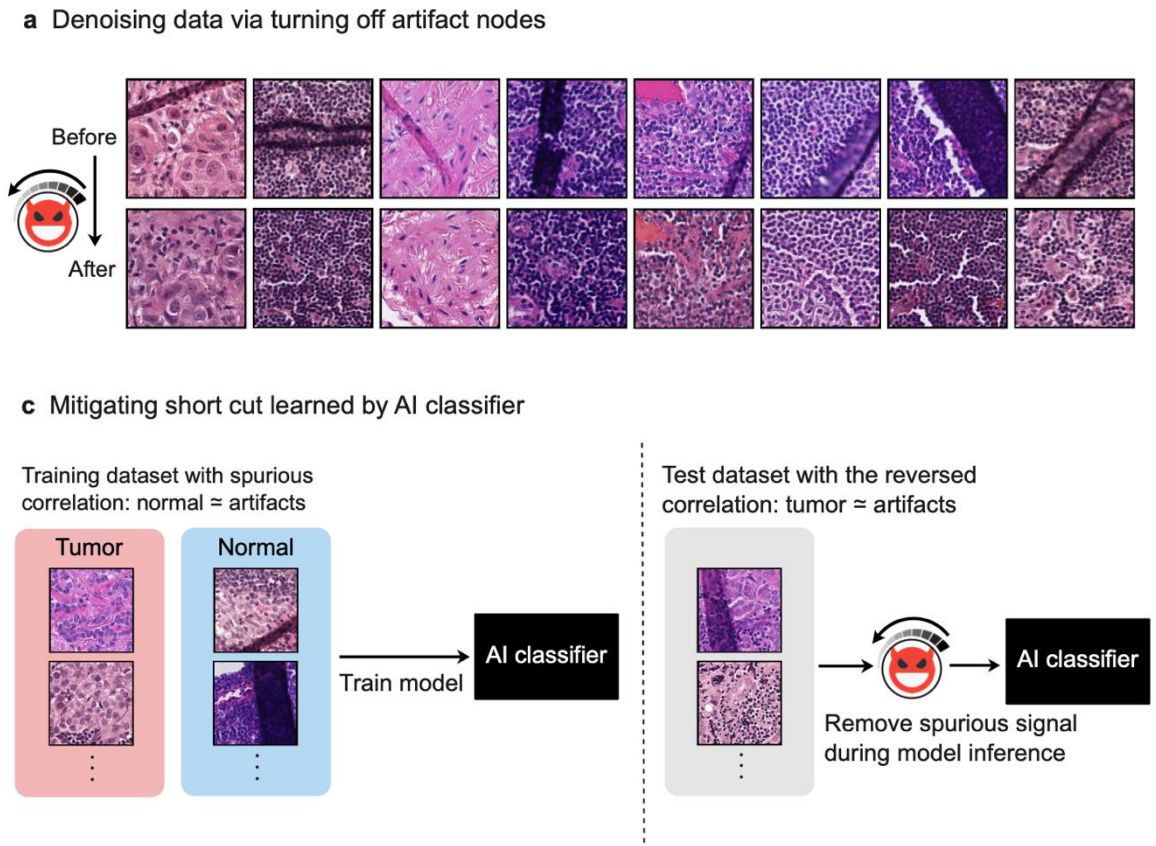
- Online version available: https://histoviewer.ksi.fit.cvut.cz

- **Histo-SAE: Zhao et al. (2026)[1]**
  - Compositional tissue representations
  - Token-level localization and expert validation



[1]Zhao, Z., Maliga, Z., et al. (2026). *Compositional and interpretable representation of histology using AI foundation models and sparse autoencoders. bioRxiv.* https://doi.org/10.64898/2026.06.03.725182

- **PICASSO: Kim et al. (2026)[1]**
  - Concept interventions and model auditing

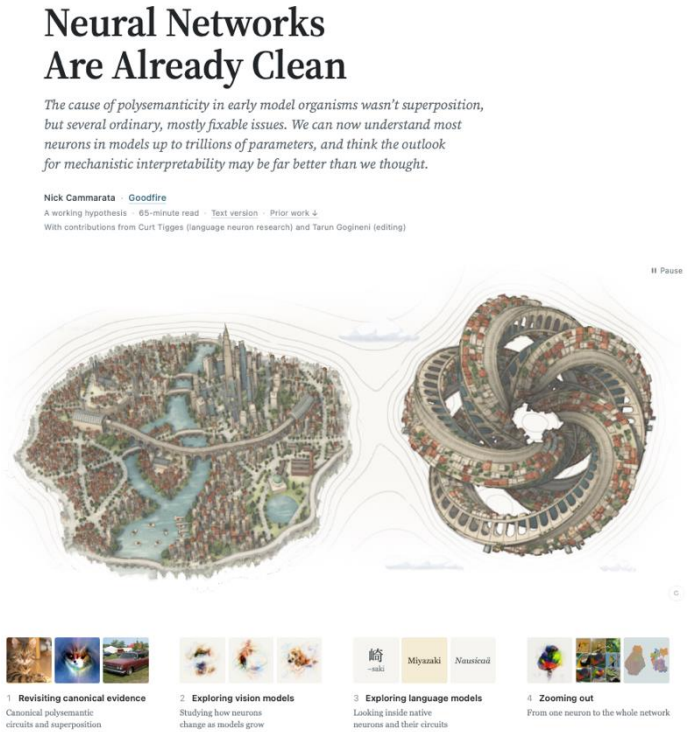


[1]Kim, C., Kaczmarzyk, J., et al. (2026). *Dissecting and directing pathology foundation models. bioRxiv. https://doi.org/10.64898/2026.06.12.731496*

But.. What if neural networks are already quite interpretable?

- **Cammarata (2026)[1] — "Neural Networks Are Already Clean"**
  - Polysemanticity does not necessarily imply superposition
  - Neurons may represent meaningful patterns we fail to recognize
  - SAEs may reorganize or fragment existing features rather than discover new ones

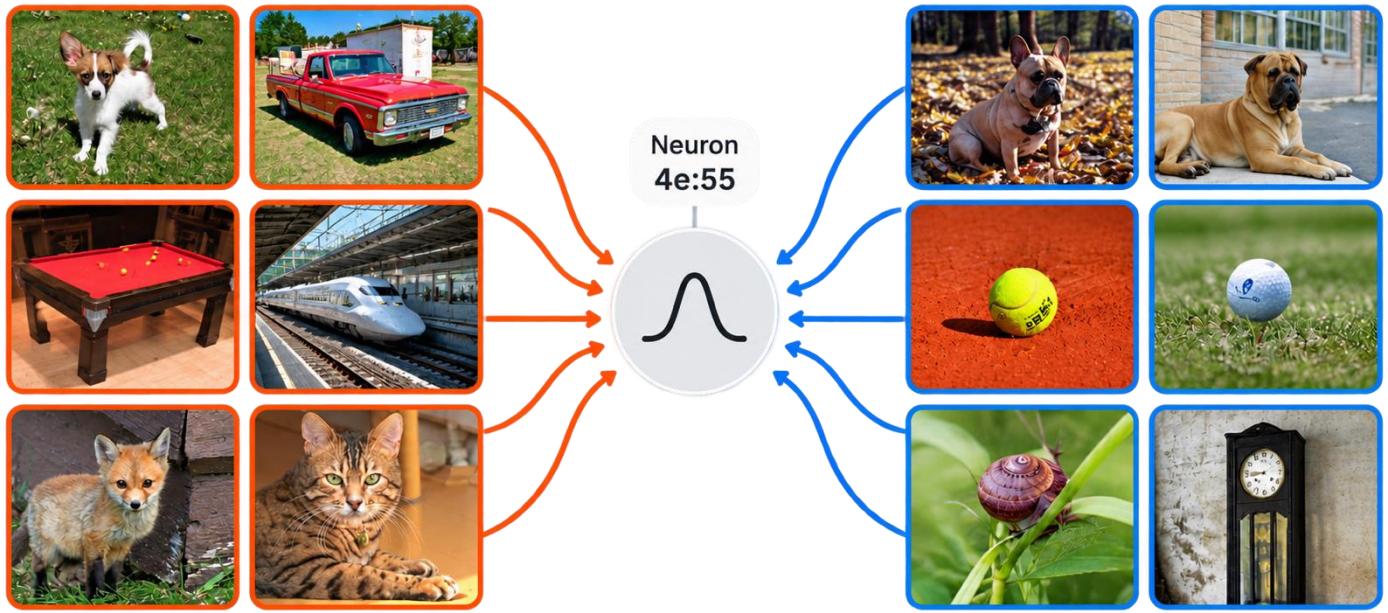
- **Question:**
  - Are SAEs revealing hidden concepts or constructing alternative descriptions?



[1]Cammarata, N. (2026). *Neural Networks Are Already Clean. Goodfire.* https://already-clean.goodfire.com

But.. What if neural networks are already quite interpretable?

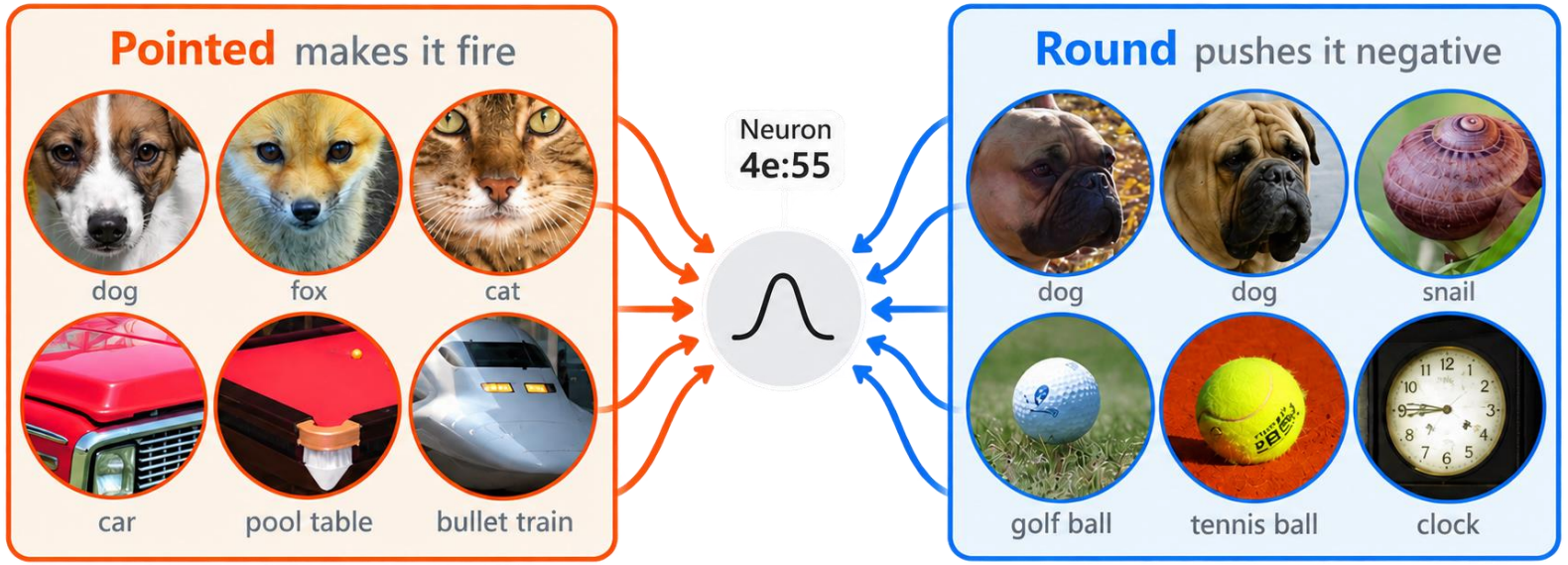
- **Cammarata (2026)[1] — "Neural Networks Are Already Clean"**
  - Polysemanticity does not necessarily imply superposition



[1]Cammarata, N. (2026). *Neural Networks Are Already Clean. Goodfire.* https://already-clean.goodfire.com

But.. What if neural networks are already quite interpretable?

- **Cammarata (2026)[1] — "Neural Networks Are Already Clean"**
  - Polysemanticity does not necessarily imply superposition



[1]Cammarata, N. (2026). *Neural Networks Are Already Clean. Goodfire.* https://already-clean.goodfire.com

Research Questions and Directions

- When can we trust a feature as a meaningful concept?
- How to evaluate such methods?

- Universality: *do features repeat across datasets and foundation models?*
- Faithfulness: *do features causally affect predictions?*
- Robustness: *do features survive distribution shift?*

- Are SAEs necessary?
  - Do they reveal meaningful structure beyond raw embeddings and internal neurons?

- **Goal:** Move from plausible explanations toward *validated understanding* of pathology foundation models.