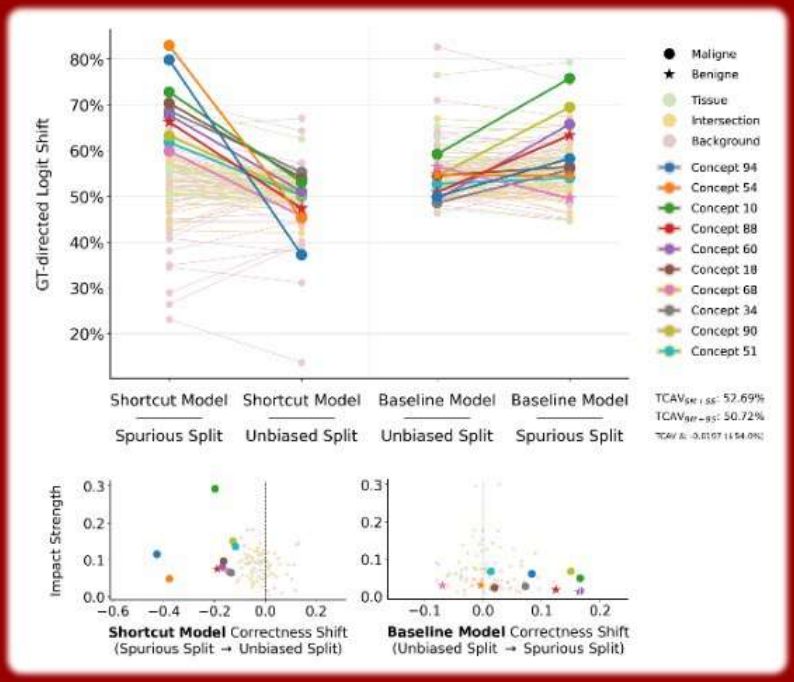


Models can retain shortcut cues — and still gate them off

Representation Is Not Reliance: Concept-Based Causal Diagnostics of Shortcut Learning in Vision

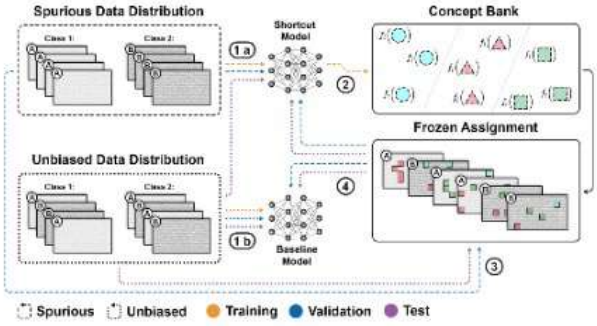
Rene Gröbner, Dr. Stefan Arnold, Dr. Tobias Clement

1 Shortcuts can survive; use is conditional

- Vision models are prone to exploiting **task-irrelevant** visual **cues** that **break under** distribution **shift.**
- Shortcut **amplification turns** these hidden **cues into reusable concepts** with trackable internal patterns.
- Concepts **can remain encoded** across models, while their **causal role depends** on the data distribution.

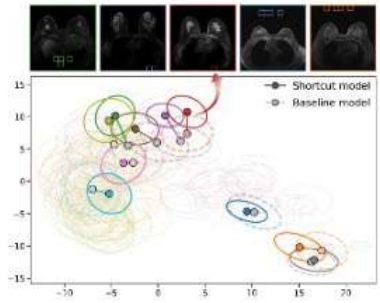
2 Diagnostic Setup

- Train one model under shortcut and one under unbiased conditions.
- Extract shortcut-amplified concepts and keep their definitions fixed.
- Test whether they transfer / drive prediction / weaken under removal.



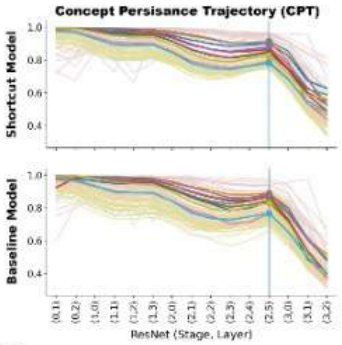
3 Concepts Remain Traceable & Stable

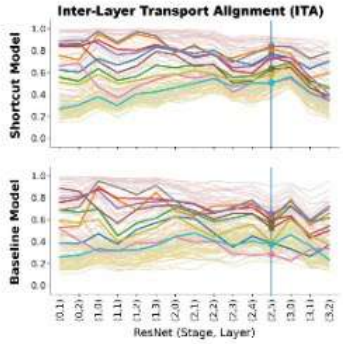
- Concepts remain recoverable: $CAV^{s} = 98.47\%$, $CAV^{b} = 97.75\%$.
- Centroid drift stays below within-cluster scale, indicating concept identity preservation.



4 Concepts Stay Coherent Across Layers

- CPT measures whether fixed concept instances remain directionally cohesive as clusters.
- High CPT indicates stable angular concept profiles across most layer transitions.
- CPT declines in the final stage as spatial detail compresses into global features.





- ITA measures whether concept instances share consistent displacement across layers.
- Coherent ITA suggests coordinated transport rather than idiosyncratic reorganization.
- Distinct trajectories suggest stable background cues and deeper tissue reorganization.

5 Limitations and Discussion

- Current evidence supports transfer, not universal applicability.
- Concept discovery depends on the selected layer and captured shortcut carriers.
- Targeted removal shows diagnostic value but remains a proof of concept.

Mechanistic Interpretability Workshop, ICML 2026

rene.edgar.groebner@fau.de