‘EHR foundation models organize clinical knowledge along axes distinct from ICD’

Why inspect Electronic Health Record Foundation Models (EHR-FMs)?

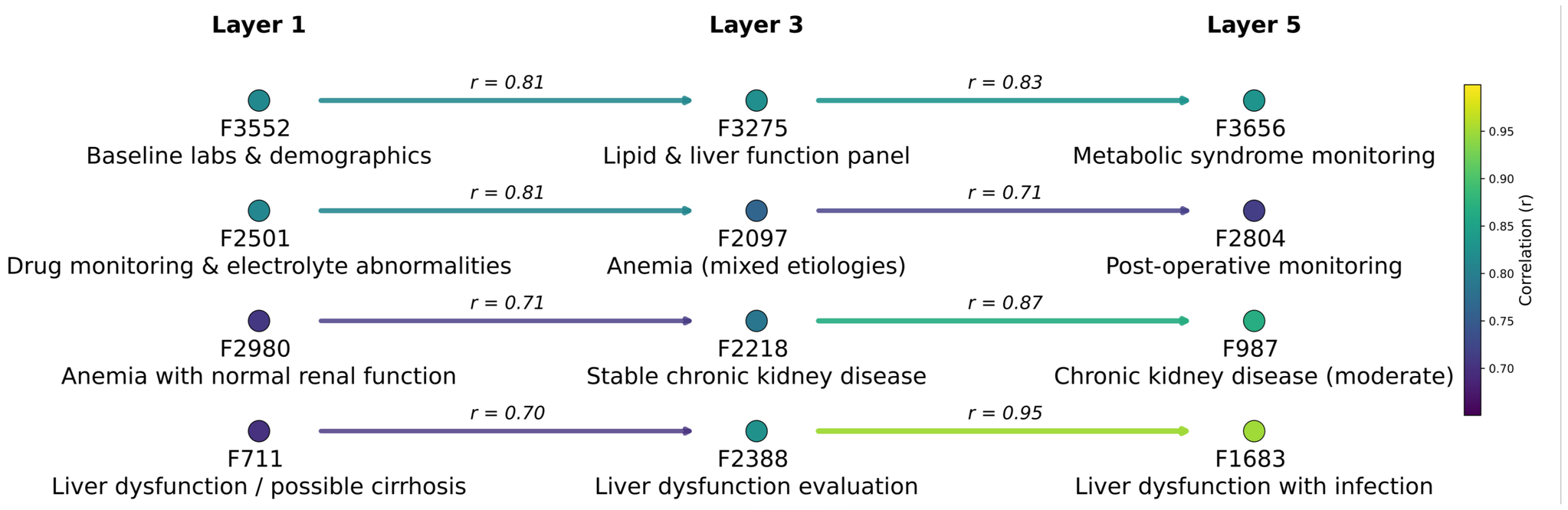
Problem: EHR foundation models are predictive but opaque — we do not know what clinical concepts they represent.

Gap: Most interpretability methods explain individual predictions rather than the model’s learned clinical ontologies.

Our approach: We train sparse autoencoders on EHR-FM residual streams to identify cross-layer clinical features.

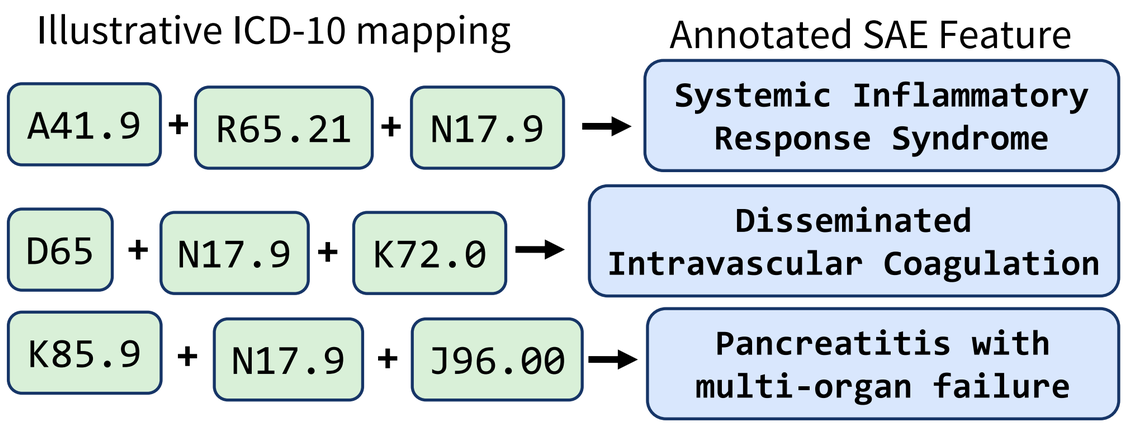
Results

Cross-layer information-flow circuits.
Each row traces a clinical concept as it composes across layers, linked by activation correlation.

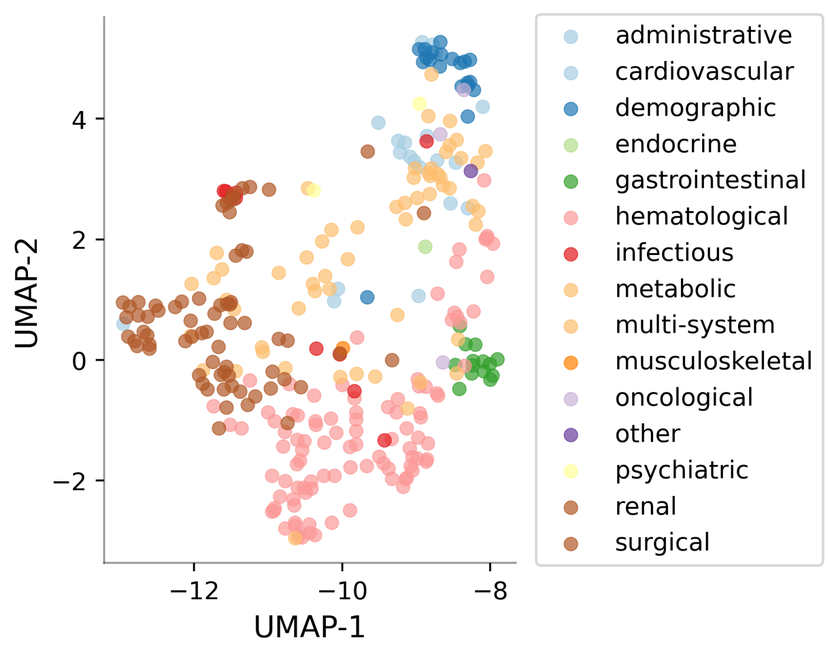


Cross-system syndromes that ICD splits across chapters collapse into a single SAE feature

26% of sampled features match to named syndromes.
Many map to ICD-10 codes spread across chapters



SAE features sharing a clinical category form coherent clusters



Methods

Train an EHR-FM on MEDS-formatted MIMIC-IV → Extract Activations and train SAEs on layers 1, 3, 5 → LLM-assisted SAE feature annotation → Validation with activation patching and comparison with ICD

6-layer transformer (27.5M params) trained on MIMIC-IV dataset v3.1 (169M events, 364,627 patients).

Top-K SAEs trained on residual-stream activations at Layers 1, 3, 5 (dictionary size 4,096, k=32).

SAE features annotated from top-activating patient contexts and validated by activation patching / ICD comparison.

Limitations

Single-site data from MIMIC-IV; SAE annotations cover only a sampled subset of the full dictionary.

SAE feature labels are prompt- and judge-LLM-dependent, and feature portability across models remains open.

Paper

Code

Probing Clinical Concepts in an EHR Foundation Model via Sparse Autoencoders

Shashank Yadav, David M. Routman, Andrew YK Foong

Radiation Oncology AI & Data Analytics AIDA

MAYO CLINIC

Mechanistic Interpretability Workshop ICML 2026