



AI in Circle U.:
Connecting Minds,
Shaping the Future
May 27-28, 2025

Explainable AI for Unsupervised Learning Turning Raw Data into Scientific Insights

Grégoire Montavon

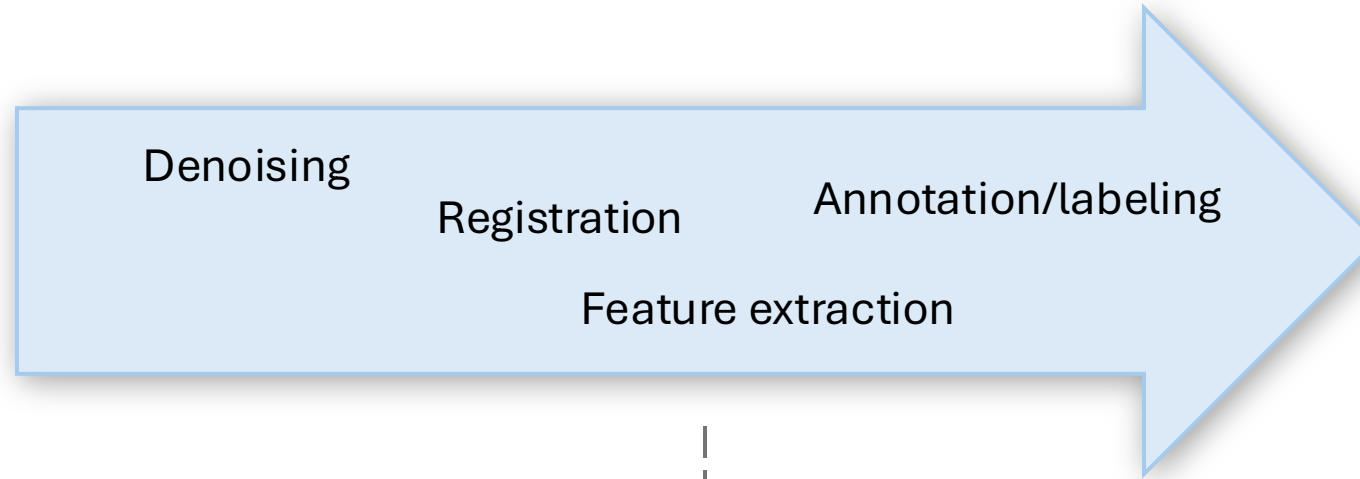


raw data



- collected automatically
- big data sets
- task-unspecific

unsupervised learning



curated data

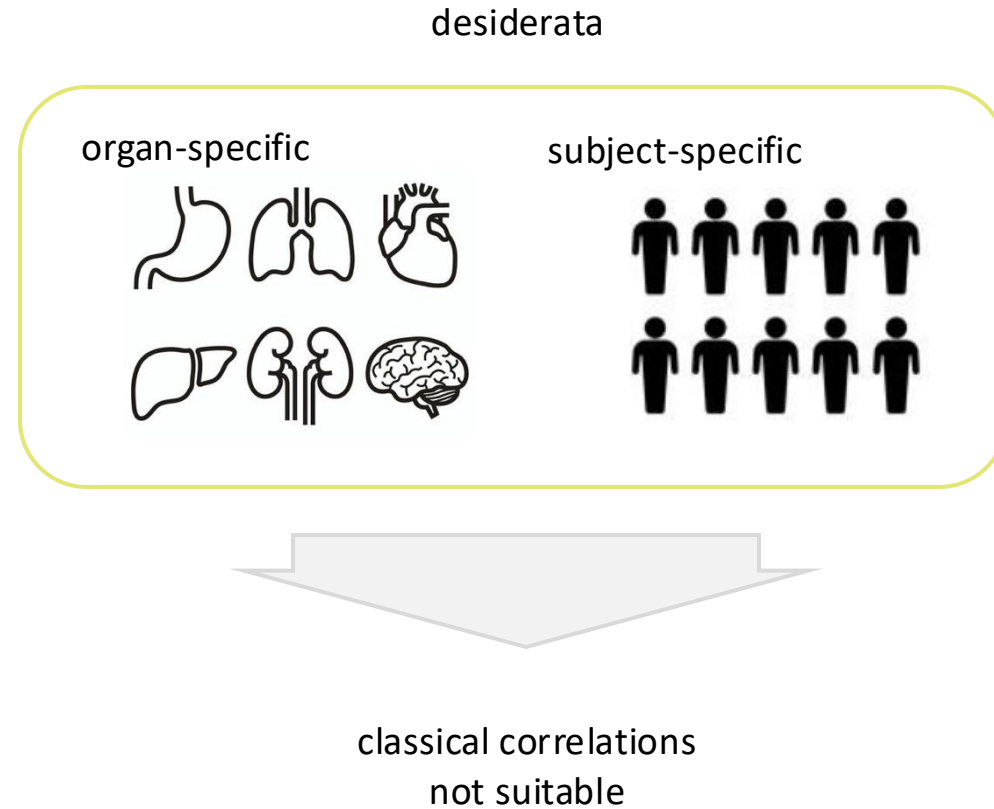
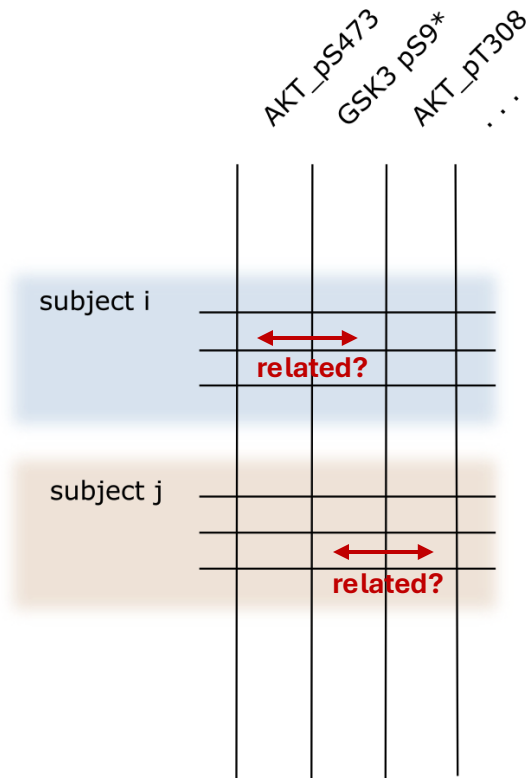


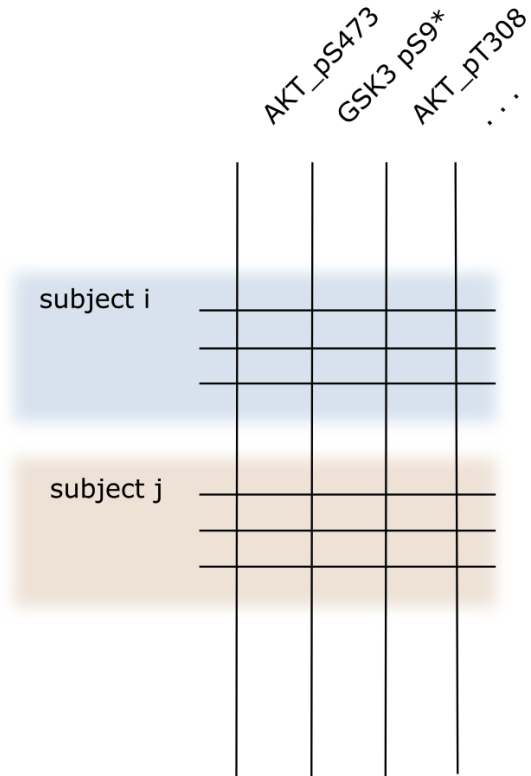
- uses human expertise
- small data sets
- task-specific

supervised learning

Inferring Regulatory Networks

Part I



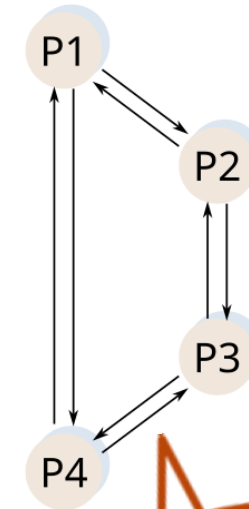


Machine learning-based

supervised:
link prediction



unsupervised:
reconstruction

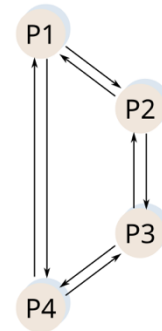
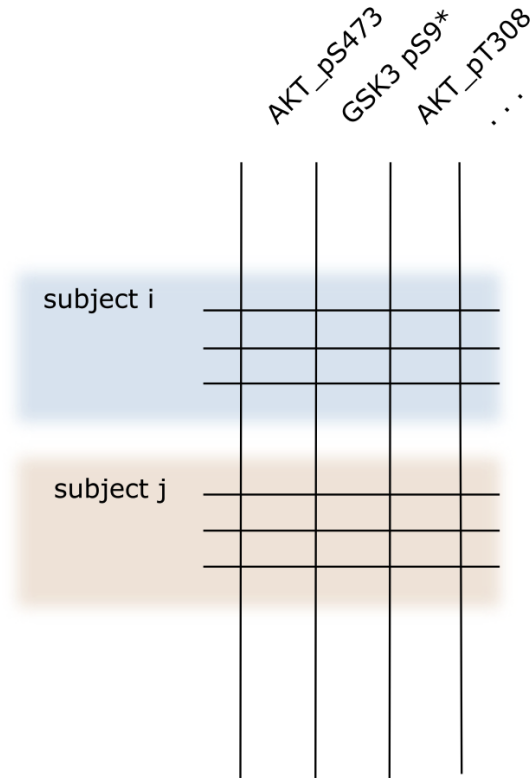


Note:

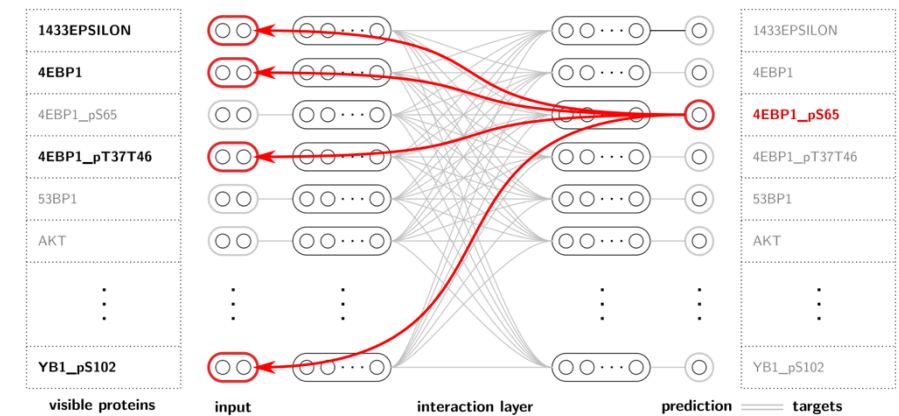
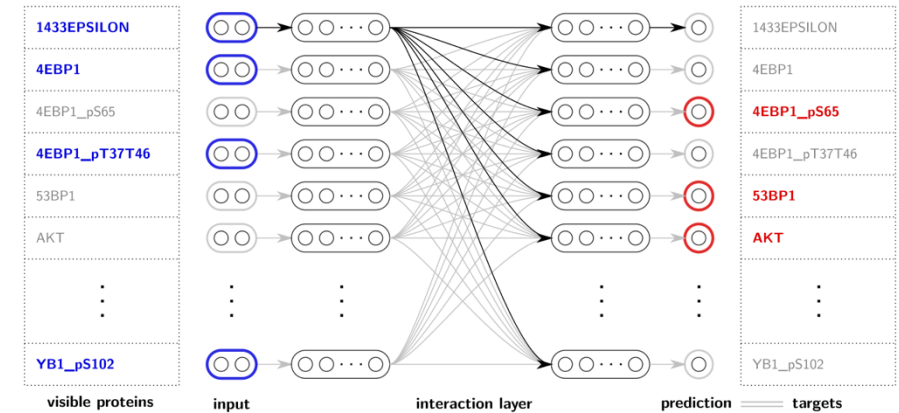
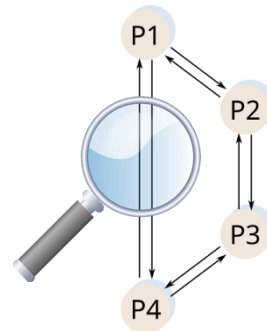
Reconstructions is only a surrogate, we want to infer **networks**

Our Unsupervised ML/XAI Approach

Step 1: Unsupervised learning

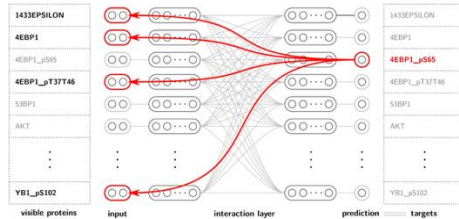


Step 2: Explainable AI

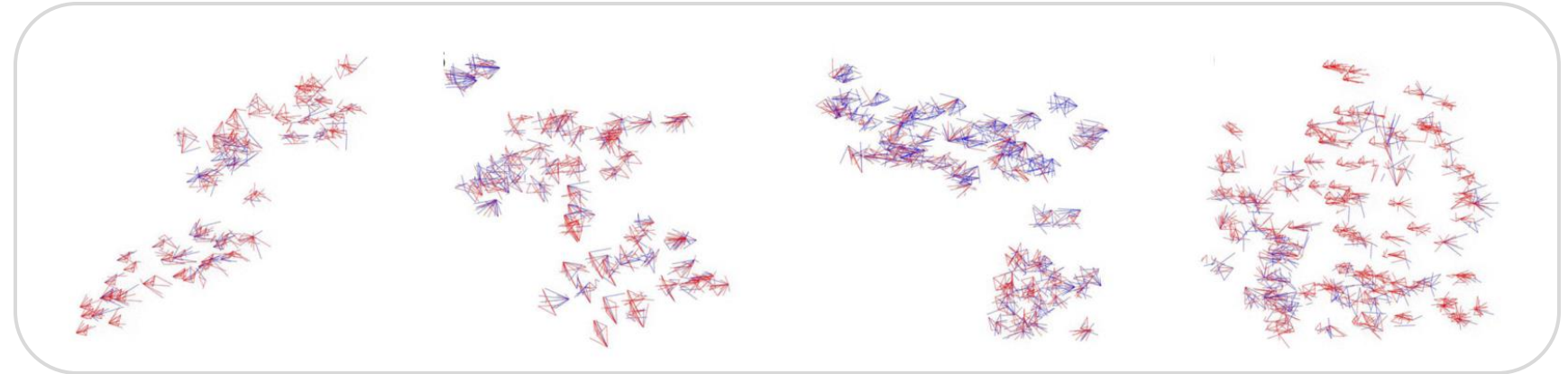


LRP (layer-wise relevance propagation)

our approach



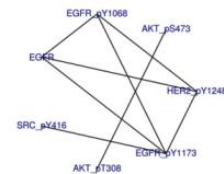
cluster aggregation



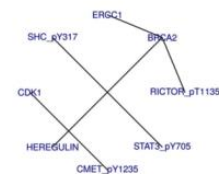
full aggregation



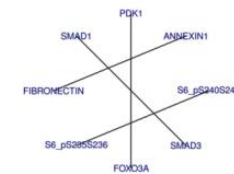
Cluster 4 (mostly GBM)



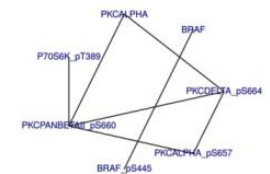
Cluster 6 (THCA)



Cluster 7 (THCA)



Cluster 4 (mostly LGG)

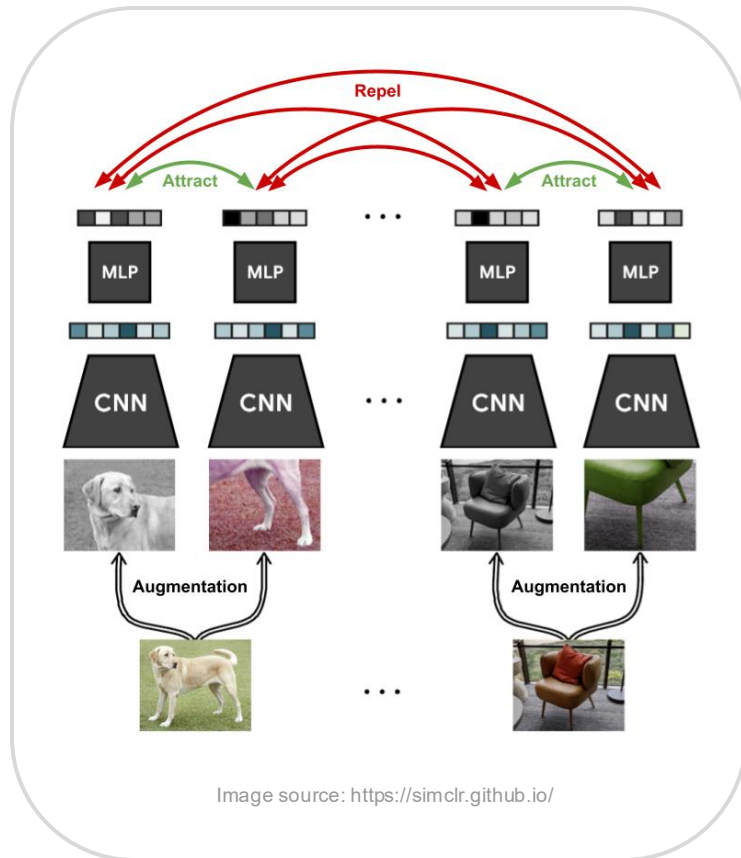


- Among the 100 strongest median interactions (out of 10,731) uncovered by our ML/XAI approach, 56 interactions were described in the Reactome database.
- In comparison, GENIE3, one of the state-of-the-art methods for network prediction, captured 42 Reactome interactions with its highest 100 predictions.

Medical Foundation Models

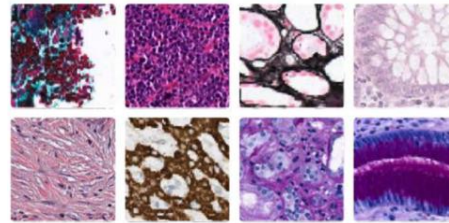
Part II

From Unsupervised Learning to Medical Foundation Models



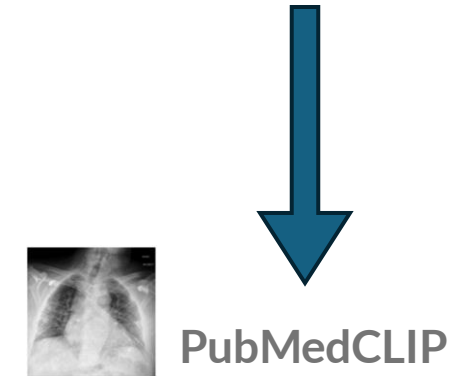
Virchow
Phikon

RudolfV



trained from scratch

OpenAI
CLIP
CONNECTING TEXT AND IMAGES



finetuned from general
foundation models

What Makes Two Images Similar?

$$\langle x, x' \rangle = \sum_{i=1}^d x_i x'_i$$

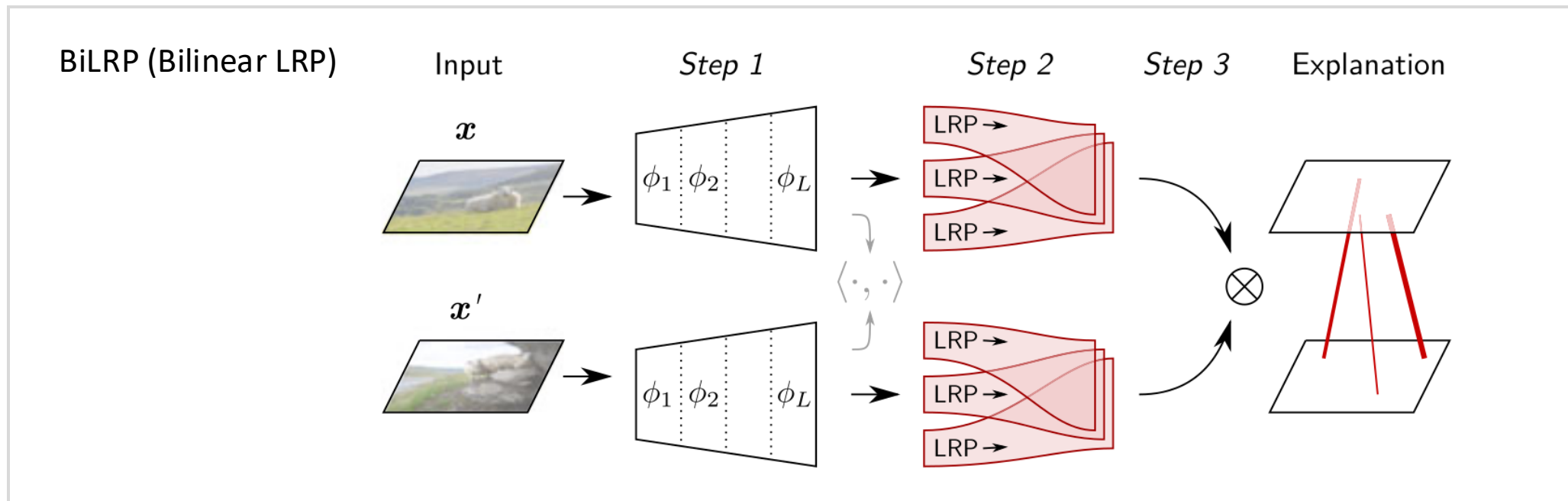
➡ attributable to input features

$$\langle Wx, Wx' \rangle = \sum_{i=1}^d \sum_{j=1}^d w_i^T w_j x_i x'_j$$

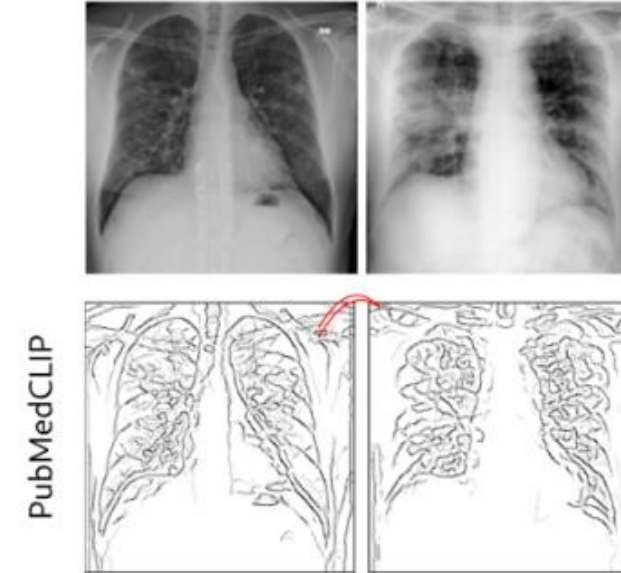
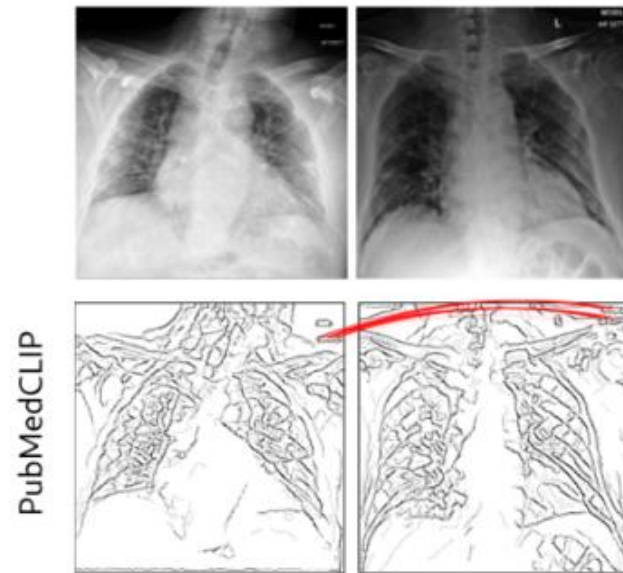
➡ attributable to pairs of input features

$$\langle \Phi(x), \Phi(x') \rangle = ?$$

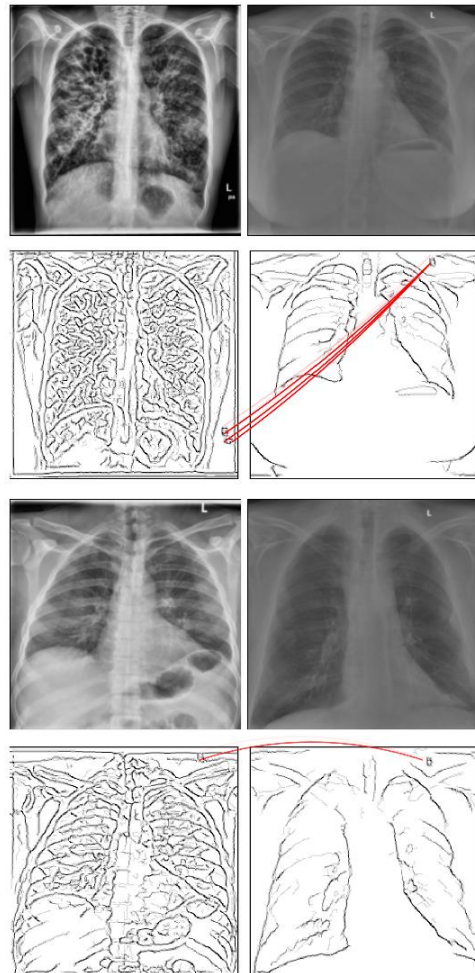
➡ specialized approach needed



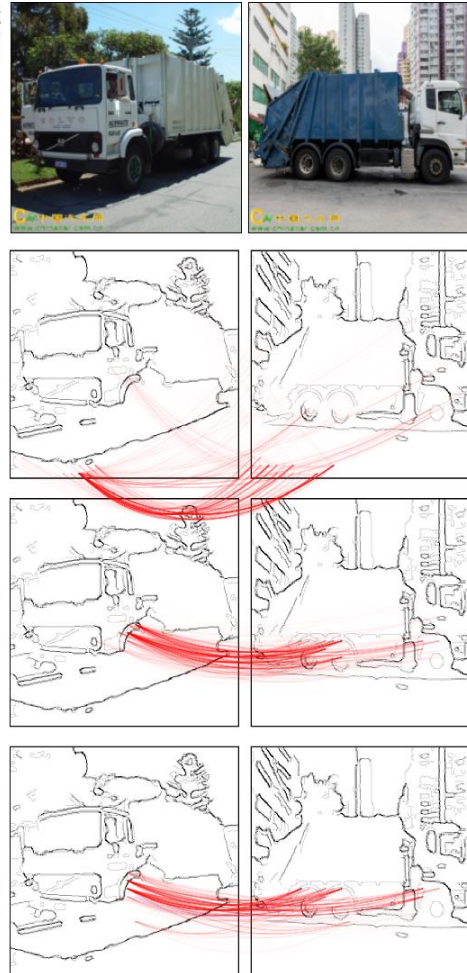
What Makes Two Images Similar?



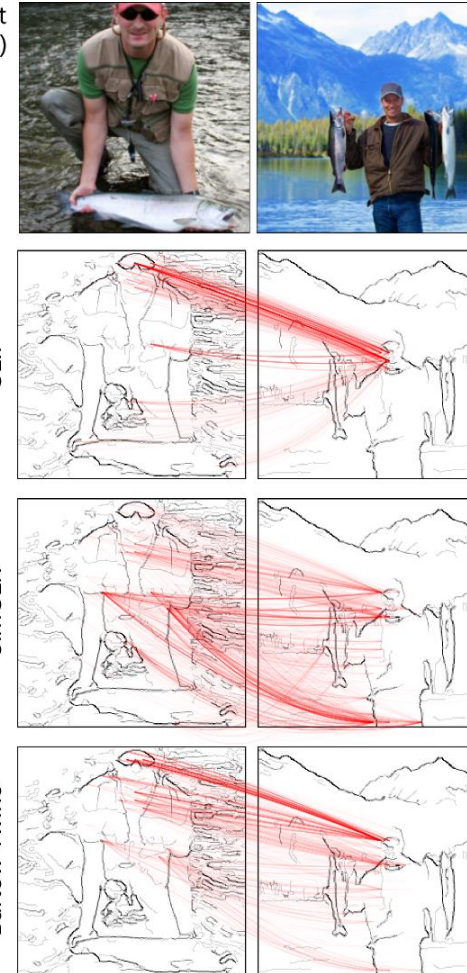
What Makes Two Images Similar?

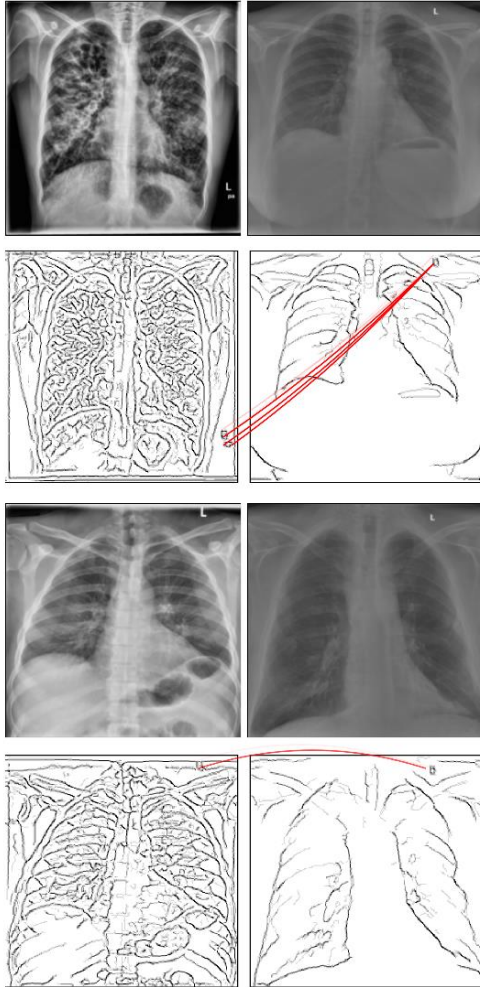


ImageNet
(gtruck-gtruck)

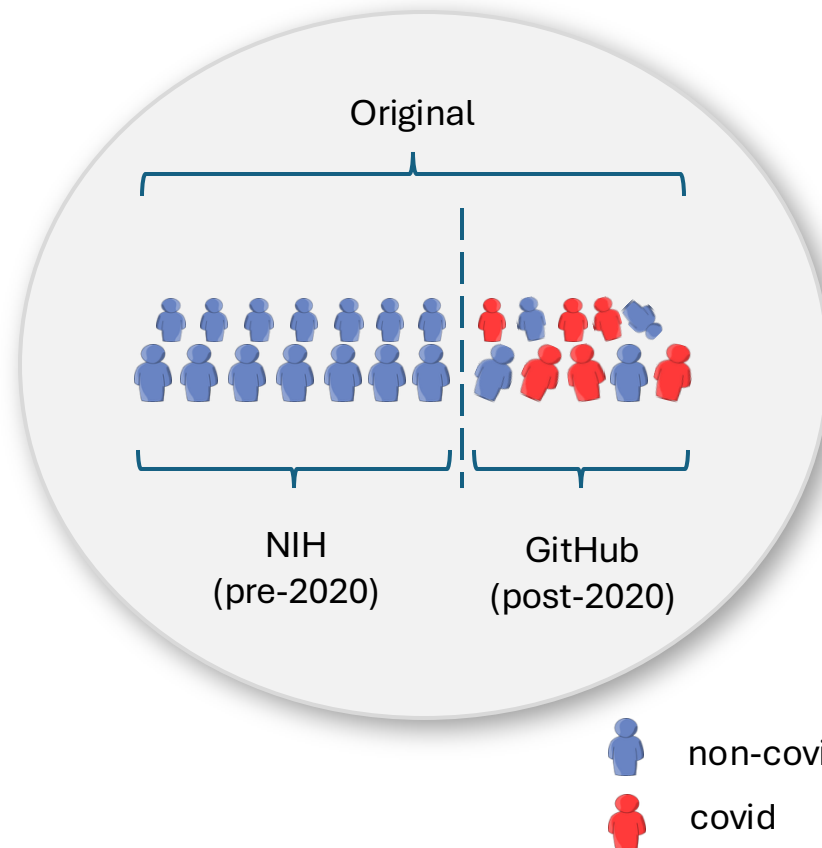


ImageNet
(coho-coho)





Downstream task (covid vs. non-covid)

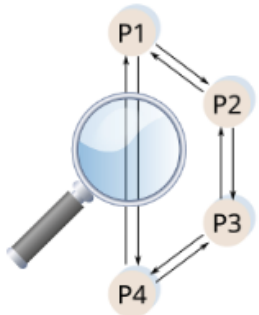


	Original	GitHub
PubMedCLIP	87.5 (FPR 18%)	81.7 ↓ (FPR 51%)

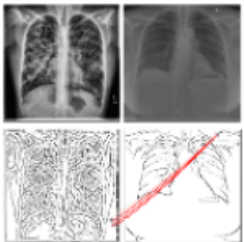
Summary



Raw data is available in abundance, not tampered by humans with a specific task in mind (more suitable for scientific **exploration**).



Joint usage of **unsupervised ML** and **Explainable AI** enables the exploration of complex data sources for extracting scientific **insights**.



Unsupervised models may **amplify/suppress** features we do not expect. Explainable AI can be used to detect these **flaws**.

References:

- P Keyl, M Bockmayr, D Heim, G Dernbach, G Montavon, KR Müller, F Klauschen. Patient-level proteomic network prediction by explainable artificial intelligence [NPJ Precision Oncology](#) 6(1):35, 2022
- P Keyl, P Bischoff, G Dernbach, M Bockmayr, R Fritz, D Horst, N Blüthgen, G Montavon, KR Müller, F Klauschen. Single-cell gene regulatory network prediction by explainable AI [Nucleic Acids Research](#), gkac1212, 2023
- O Eberle, J Büttner, F Kräutli, KR Müller, M Valleriani, G Montavon. Building and Interpreting Deep Similarity Models [IEEE Transactions on Pattern Analysis and Machine Intelligence](#), 44(3):1149-1161, 2022
- J Kauffmann, J Dippel, L Ruff, W Samek, KR Müller, G Montavon. Explainable AI reveals Clever Hans effects in unsupervised learning models, [Nature Machine Intelligence](#) 7:412-422, 2025