Winter School on Causality and Explainable AI

Uncovering Input-Target Associations with Explainable AI

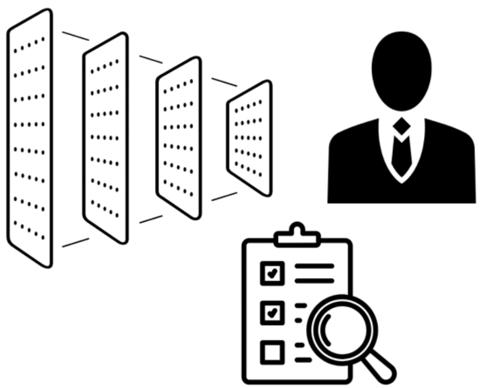
Grégoire Montavon

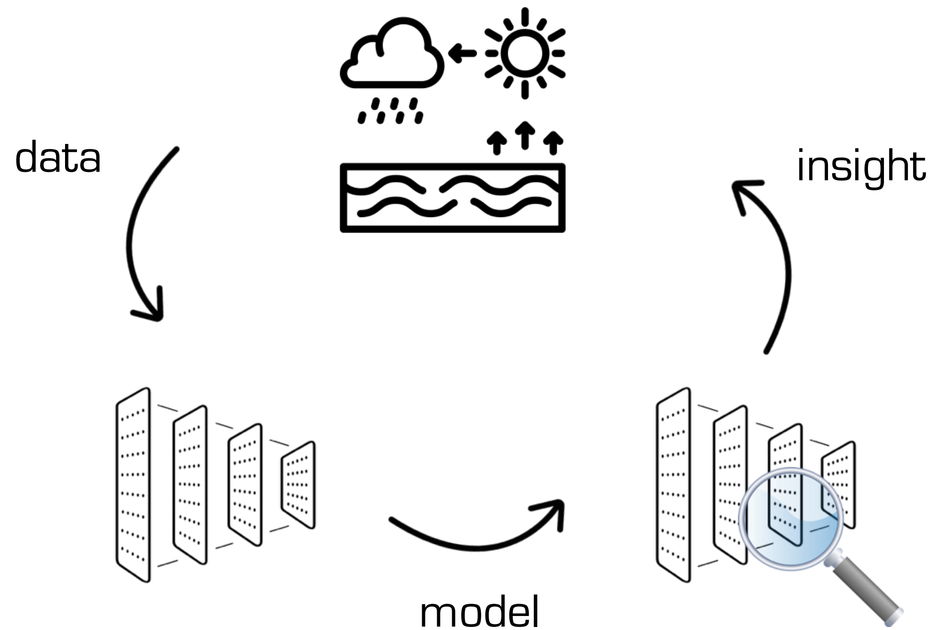
22 October 2025

CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN

BIFOLD

Two Distinct Uses of XAI





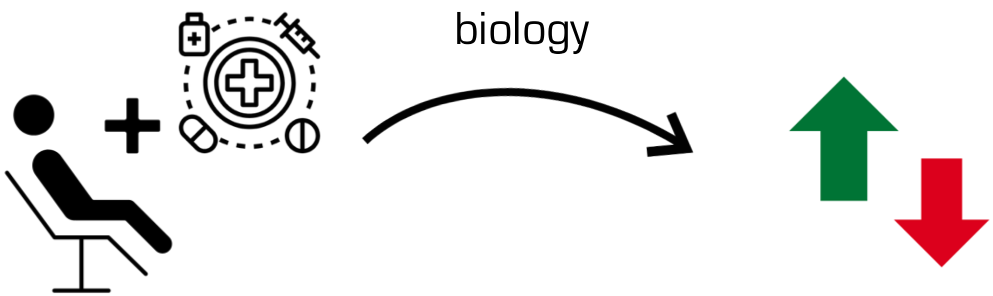
1. XAI for ML Auditing

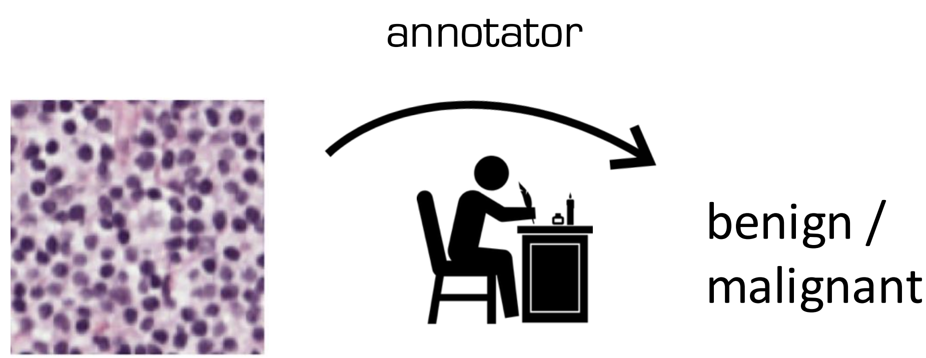
- Object of interest is the ML model. XAI is the tool.

2. Understanding

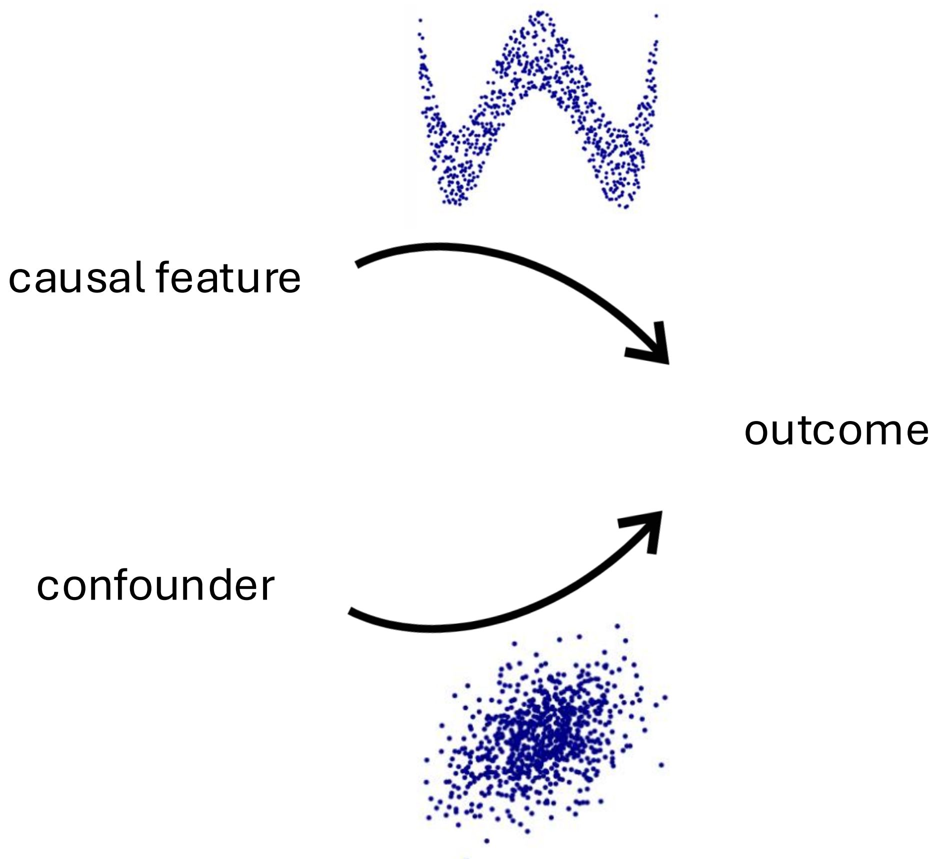
- Object of interest is the natural system or process. ML/XAI are the tools.

Examples of Systems of Interest





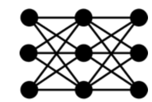
Hypothesis on Causal Features



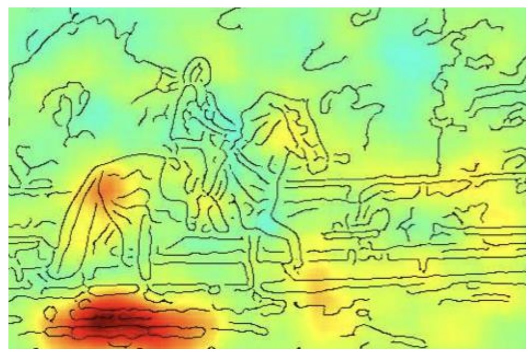
Empirical Evidence

annotation → horse

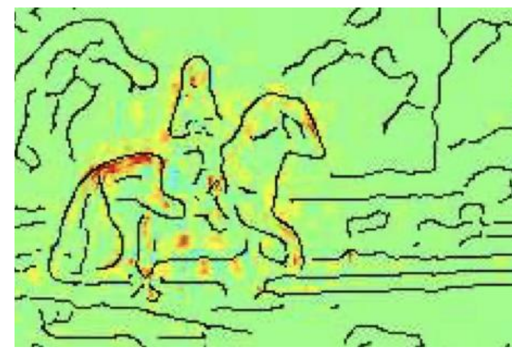
pre-2012



post-2012



- limited data
- simple models
- weak correlates
- less generality



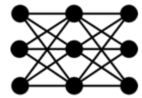
- big data
- complex models
- strong correlates
- more generality

Lapuschkin et al. CVPR'16

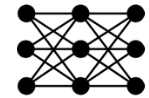
Empirical Evidence

BIFOLD

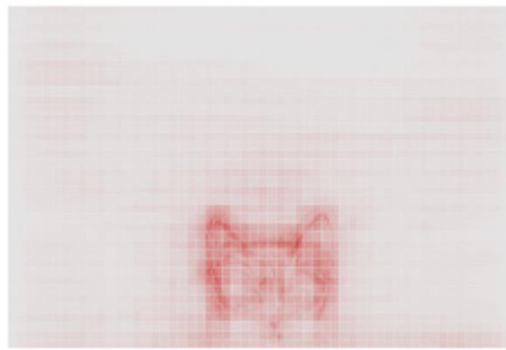
$\xrightarrow{\text{annotation}}$ cat

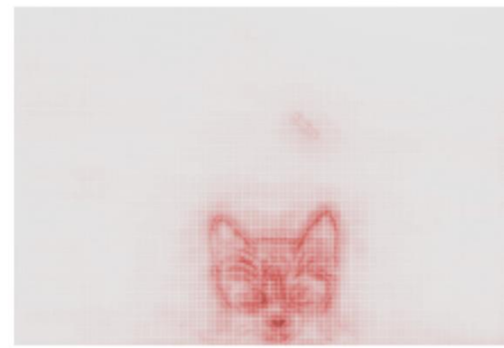


2012



2014





2012:

- limited data
- simple models
- weak correlates
- less generality

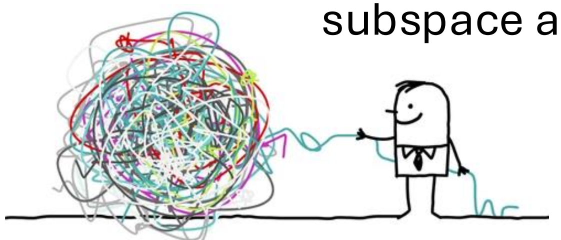
2014:

- big data
- complex models
- strong correlates
- more generality

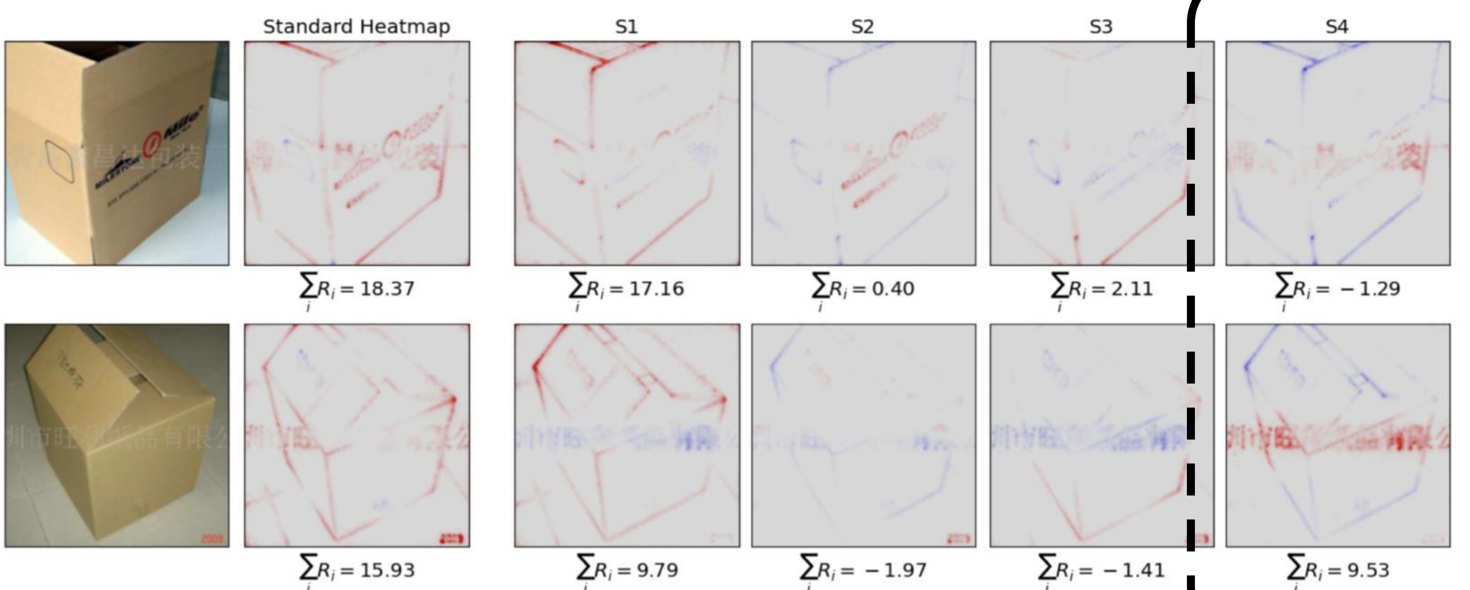
Montavon et al. Pattern Recognition (2017)

Deconfounding Methods

disentangled relevant subspace analysis (DRSA)

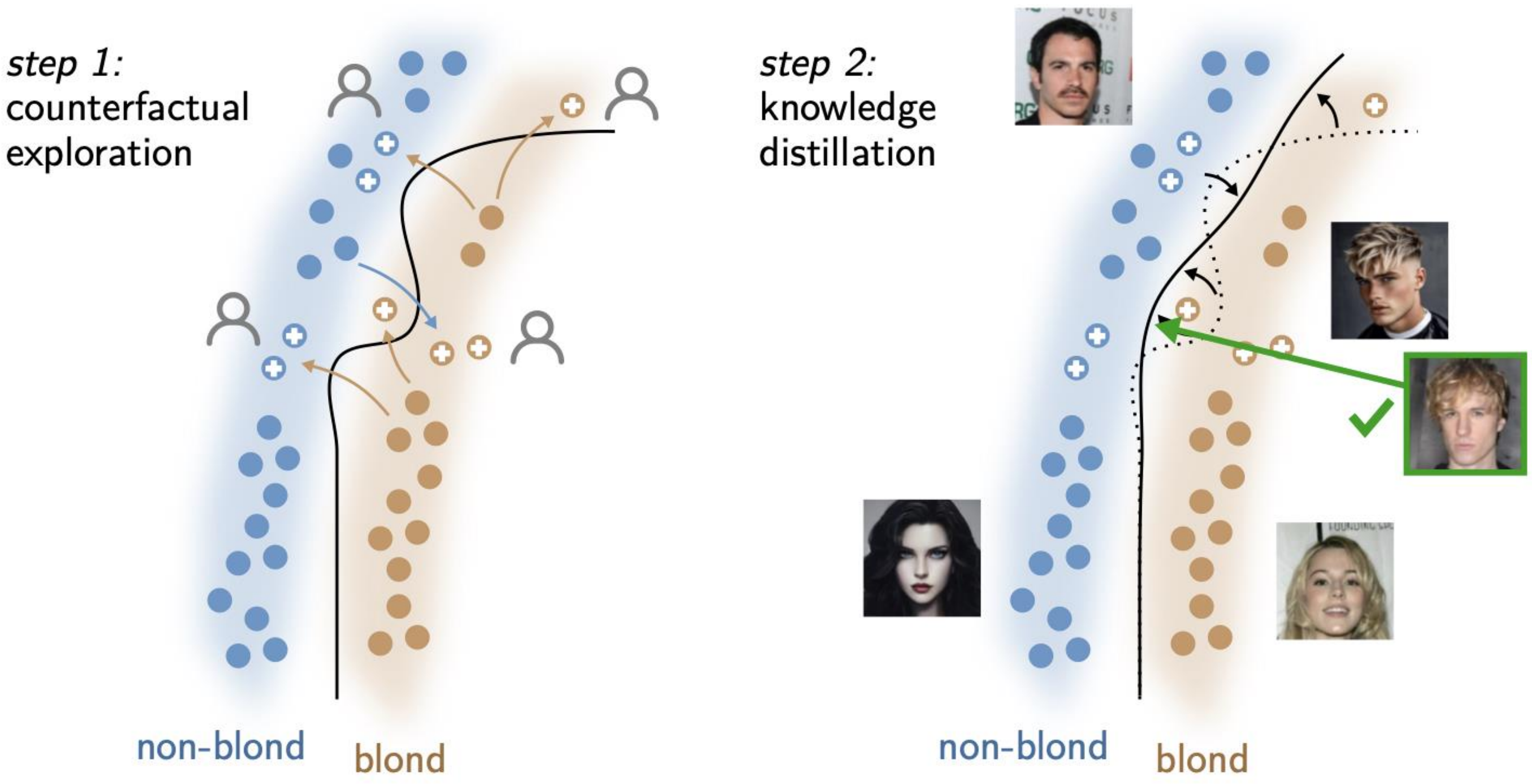


human inspection + pruning



Chormai et al. IEEE TPAMI (2024)

Deconfounding Methods



Bender et al. arXiv’25

Trends in AI for Medicine

BIFOLD

- Large datasets publicly available.
- State-of-the-art ML architectures (e.g. transformers, Mamba, etc.) being applied.
- Methods to detect/remove confounders being developed.

uk biobank

TCGA

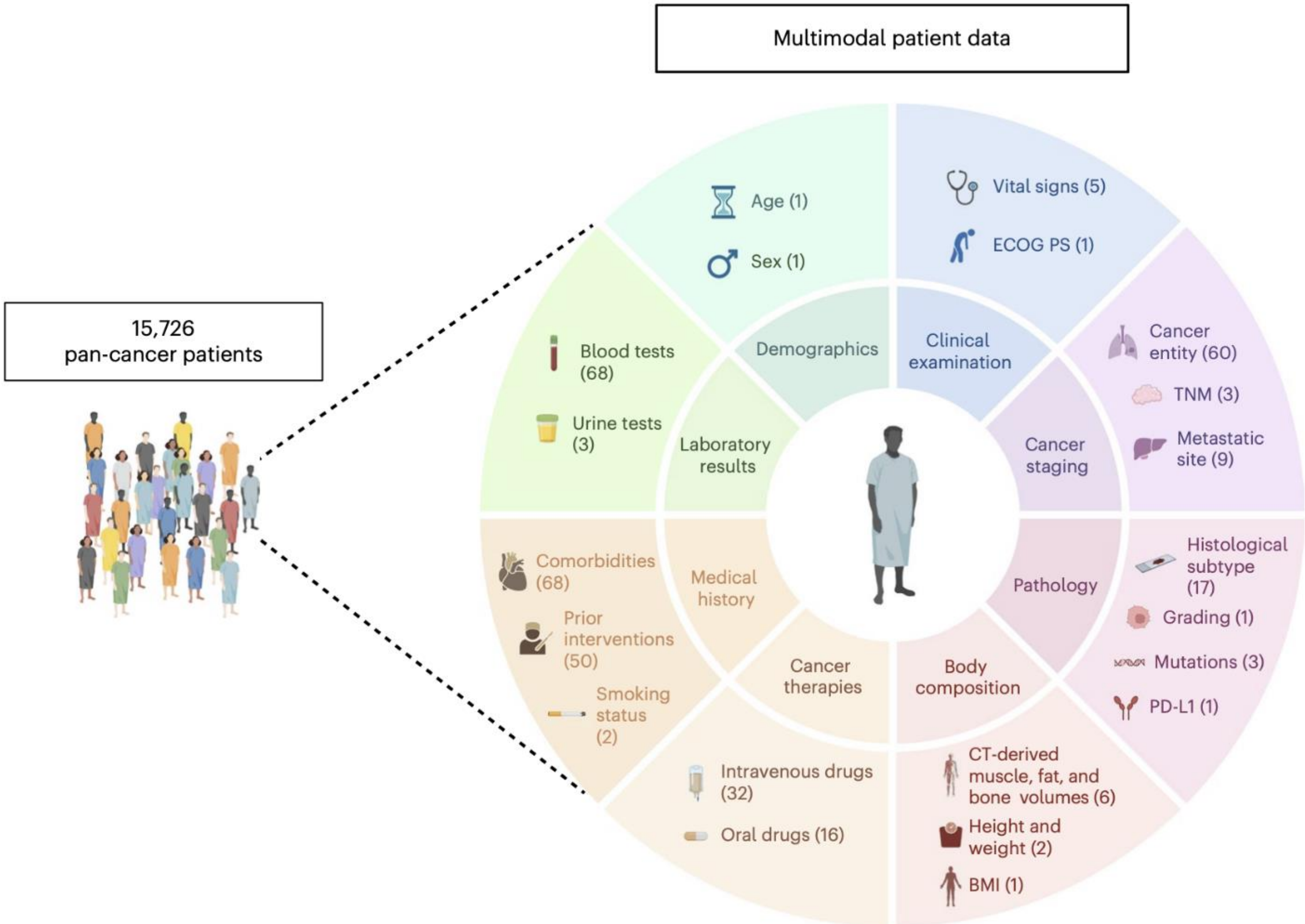
Meditron

Med-PaLM

OpenAI CLIP

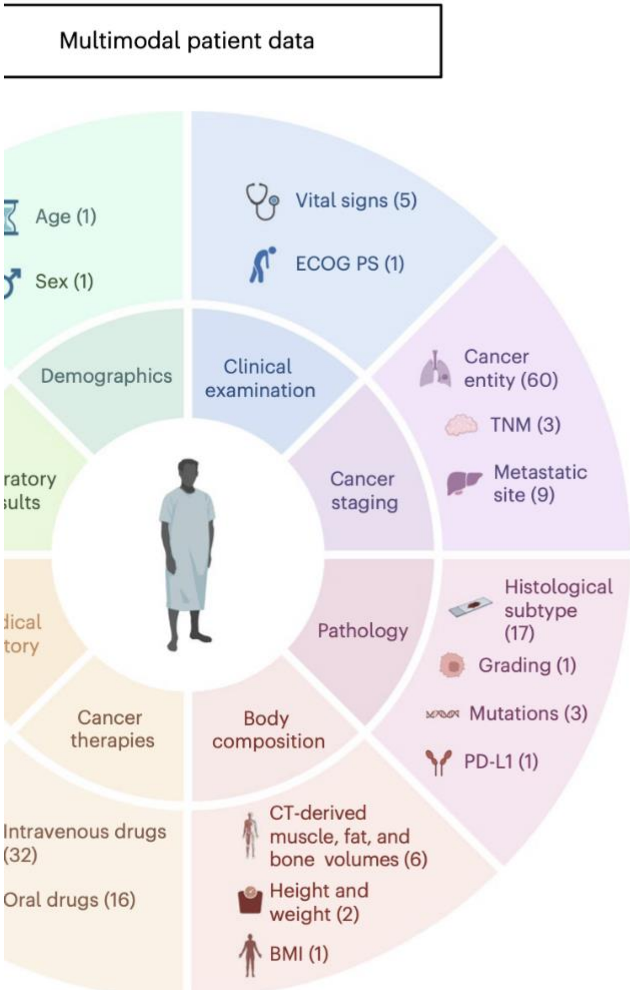
CONNECTING TEXT AND IMAGES

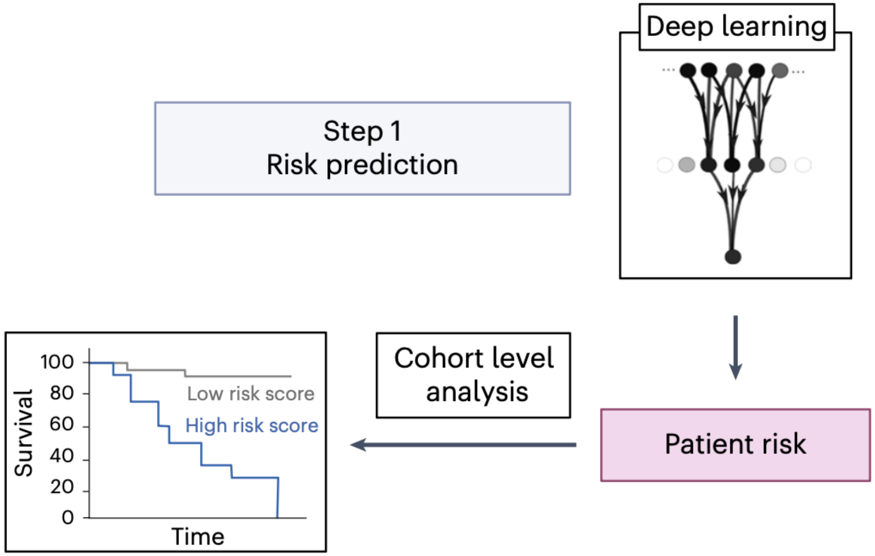
Pan-Cancer XAI Analysis



Keyl et al. Nature Cancer (2025)

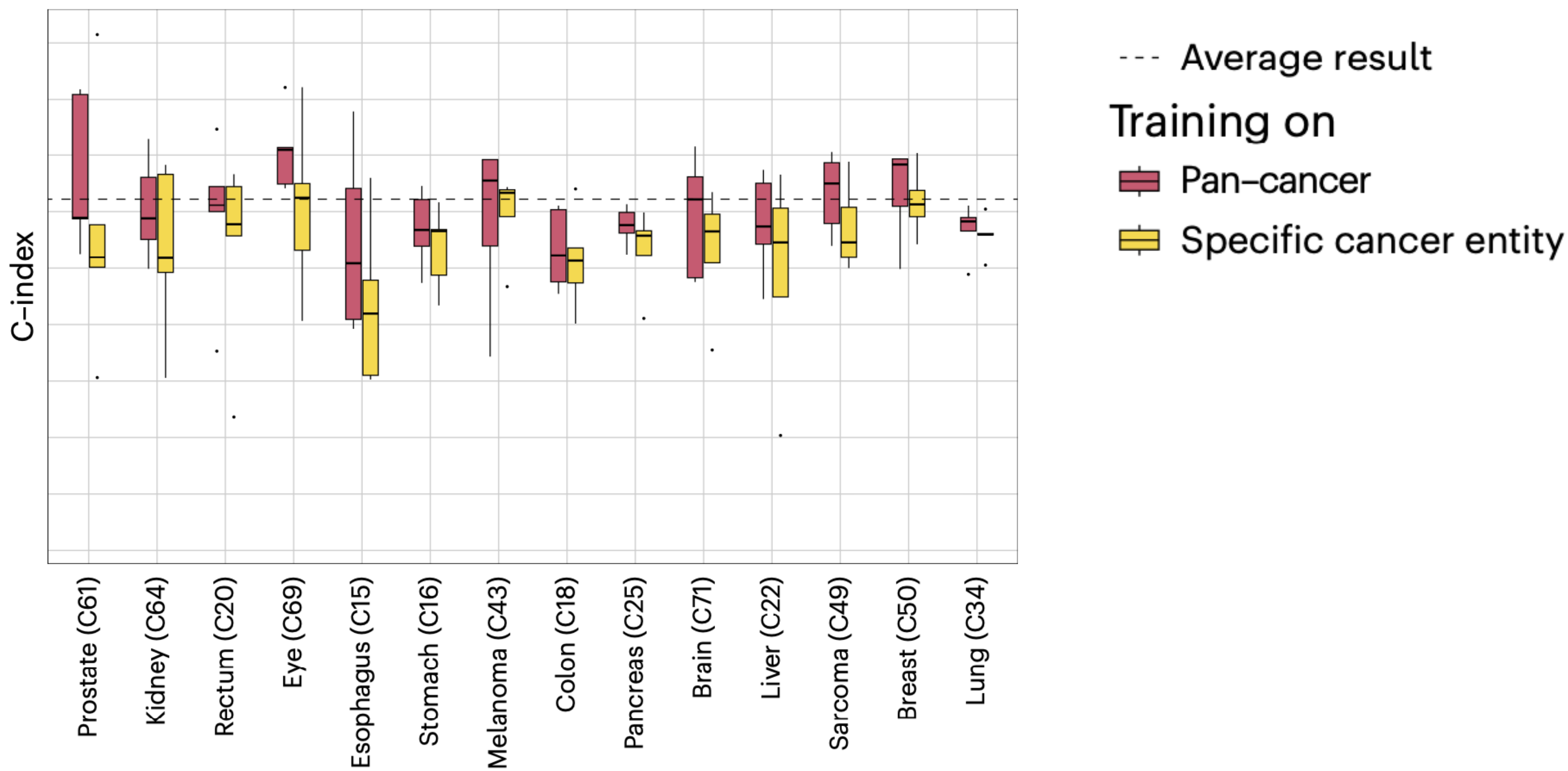
Pan-Cancer XAI Analysis





Keyl et al. Nature Cancer (2025)

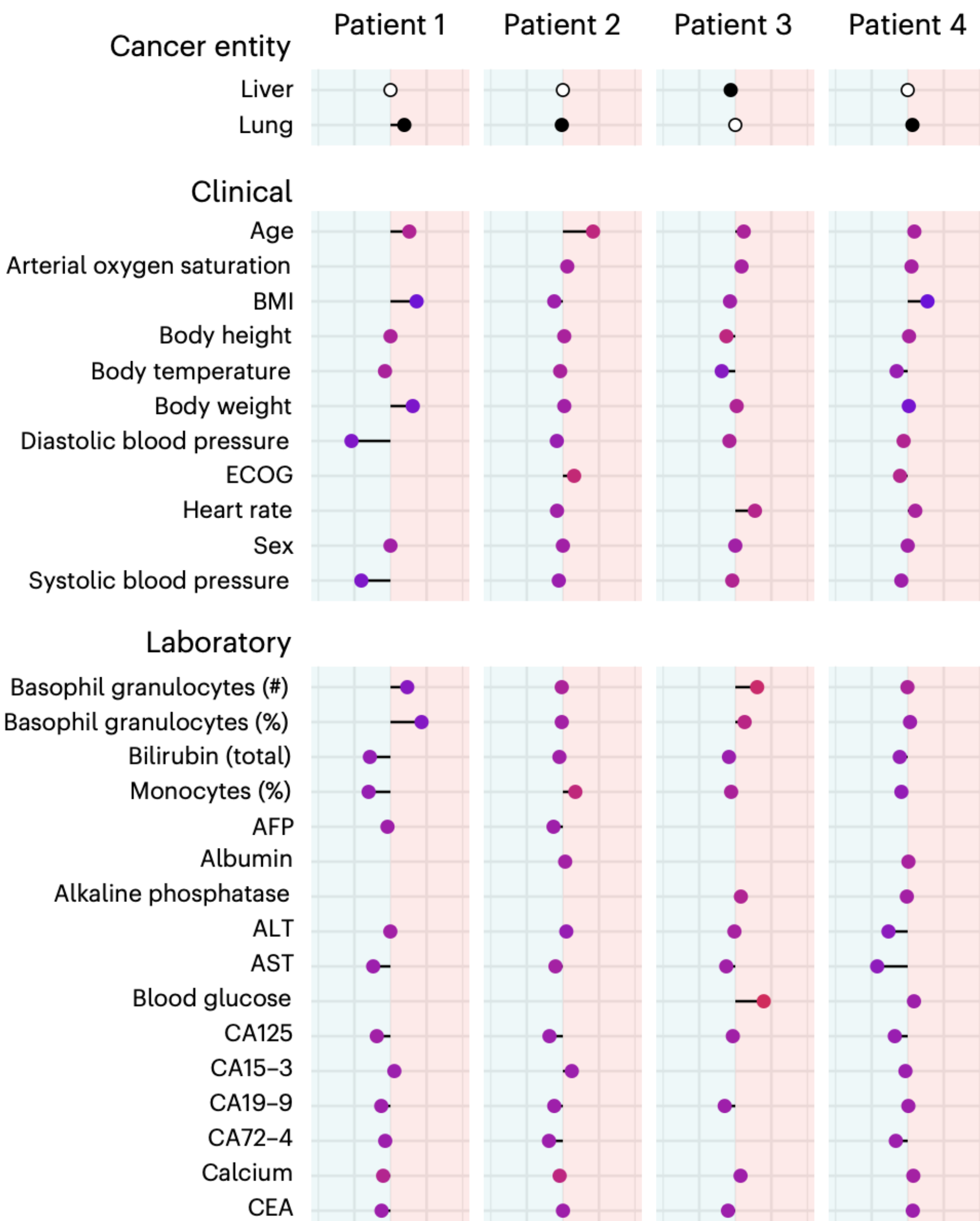
Pan-Cancer XAI Analysis

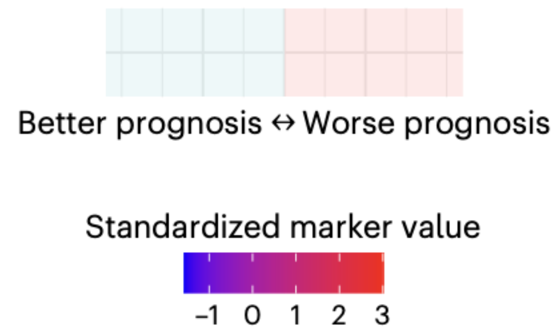


Pan-cancer increases data and benefits from statistical regularities across cancers.

Keyl et al. Nature Cancer (2025)

Pan-Cancer XAI Analysis





Deconfounding through

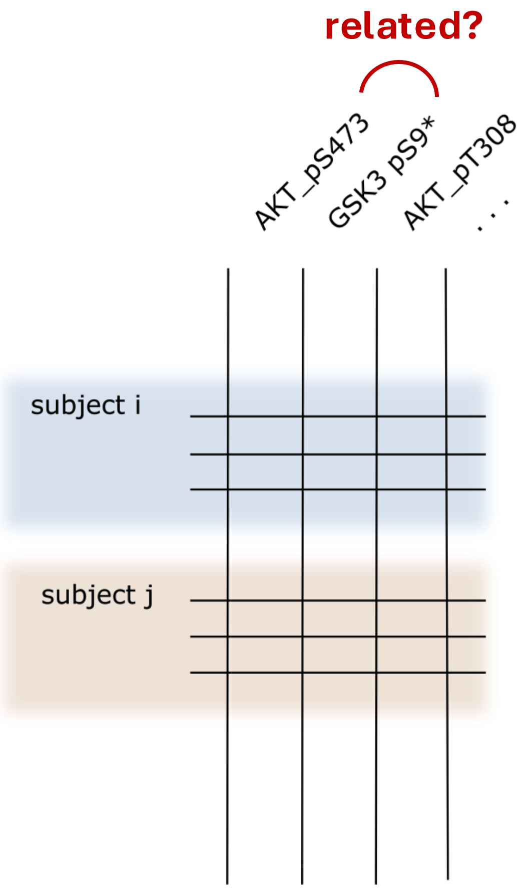
- High accuracy (large datasets + nonlinear model)
- Early stopping + input dropout

Keyl et al. Nature Cancer (2025)

Inferring Regulatory Networks

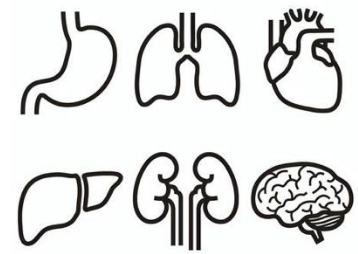
Part II

Inferring Regulatory Networks

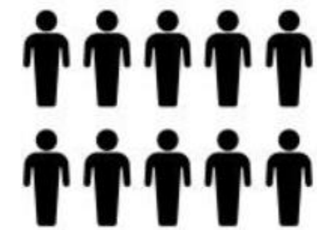


desiderata

organ-specific



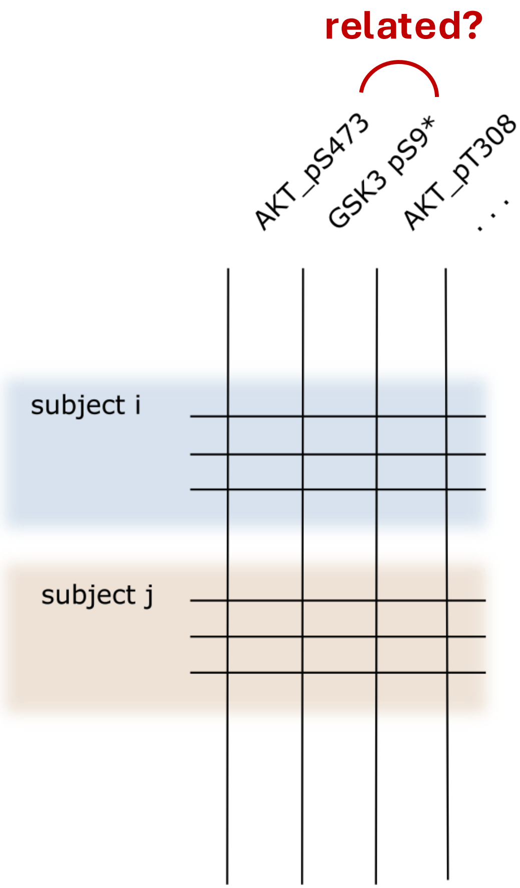
subject-specific



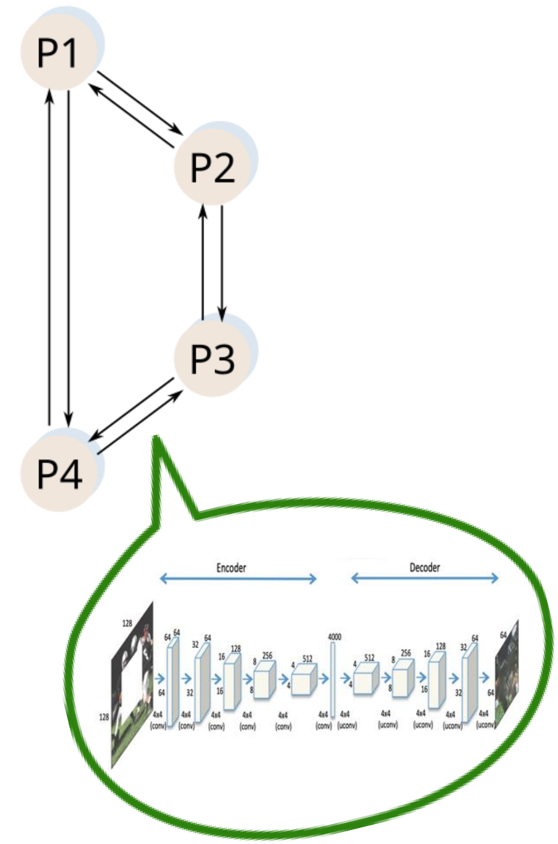
classical correlations
not suitable

Keyl et al. NPJ Precis Oncol 2022; Nucleic Acids Res 2023

Inferring Regulatory Networks



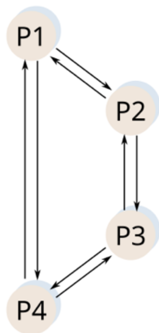
Unsupervised learning approach

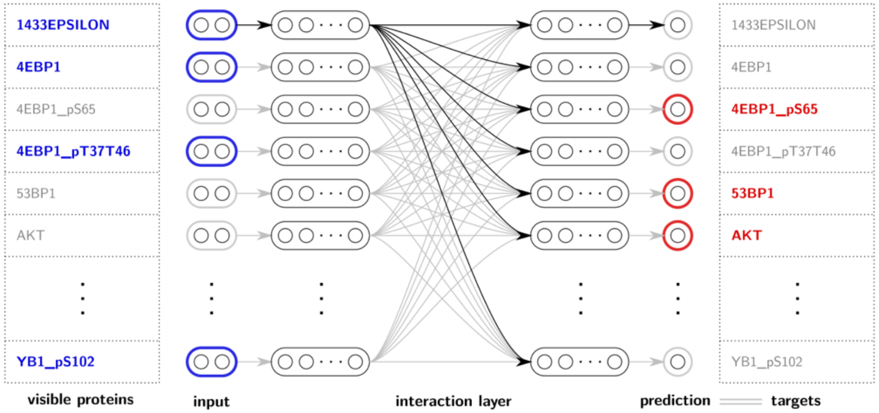


Keyl et al. NPJ Precis Oncol 2022; Nucleic Acids Res 2023

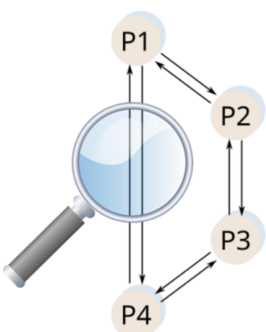
Unsupervised ML/XAI Approach

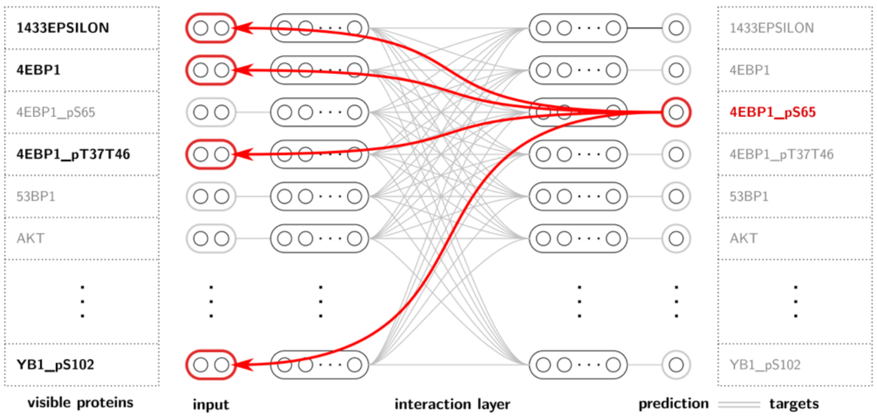
Step 1: Unsupervised learning





Step 2: Explainable AI



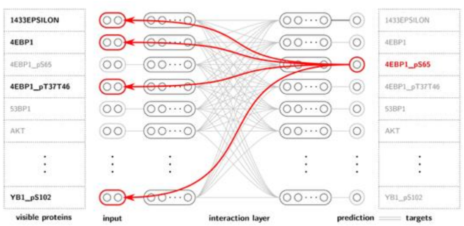


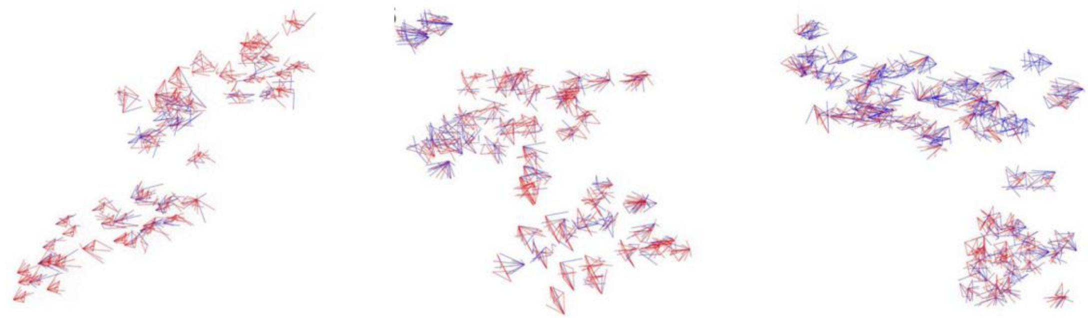
LRP (layer-wise relevance propagation)

Keyl et al. NPJ Precis Oncol 2022; Nucleic Acids Res 2023

Inferring Regulatory Networks

our approach



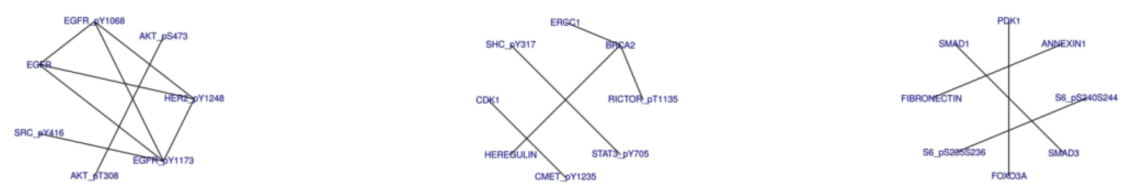


cluster aggregation

Cluster 4
(mostly GBM)

Cluster 6
(THCA)

Cluster 7
(THCA)



full aggregation

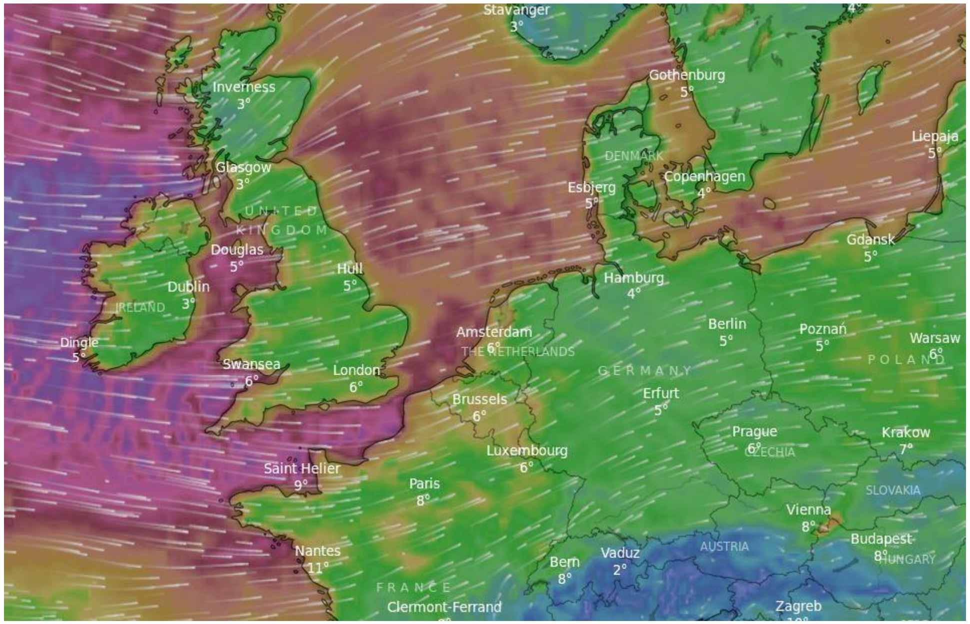
- Among the 100 strongest median interactions (out of 10,731) uncovered by our ML/XAI approach, 56 interactions were described in the Reactome database.
- In comparison, GENIE3, one of the state-of-the-art methods for network prediction, captured 42 Reactome interactions with its highest 100 predictions.

Keyl et al. NPJ Precis Oncol 2022

Input-Uncertainty Associations

Part III

Storm Sabine



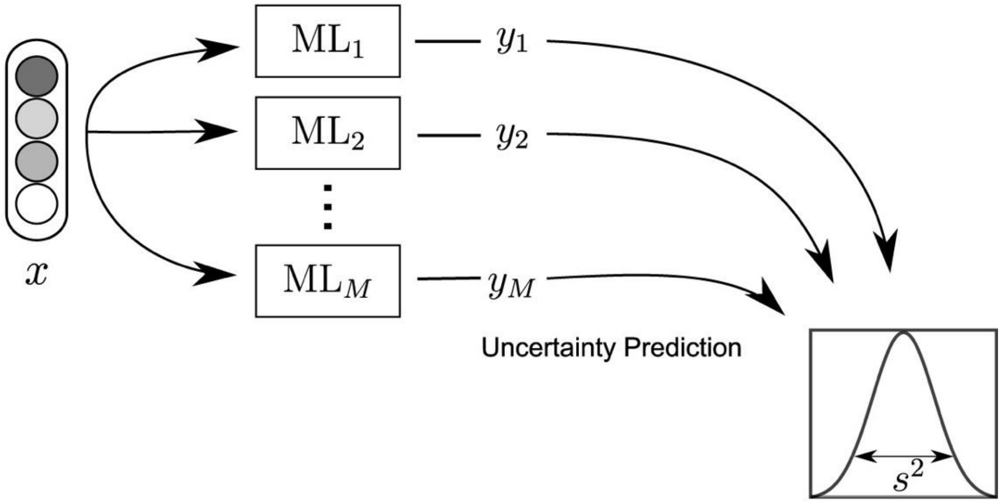
Storm Sabine
9-11 Feb 2020

- High volatility in electricity prices observed.
- What are the factors that drive price volatility?
- How can we model volatility?

Volatility as Uncertainty

Ensemble-based ML model

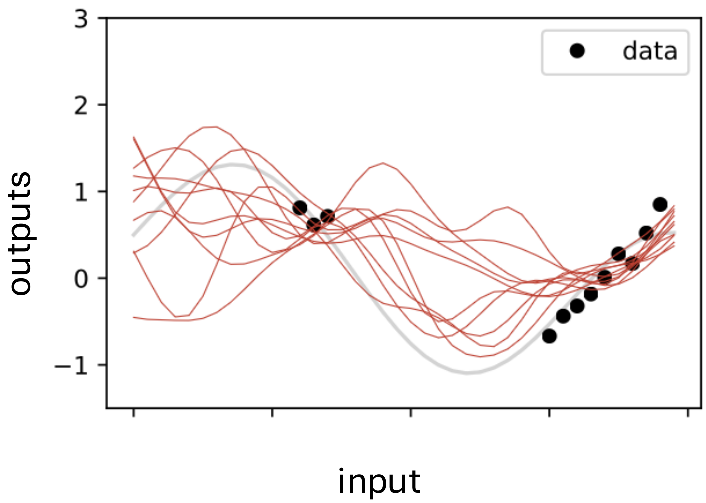
$$x \mapsto Var\{y_1, \dots, y_M\}$$



Advantages:

✓ Positive-constrained

✓ Prior encoded that uncertainty should increase in unknown situations.



Bley et al. Pattern Recognition 2025

Explaining Uncertainty

Explanations of sums

$$\mathcal{E}\left\{\sum_m \alpha_m y_m\right\} = \sum_i \alpha_m \mathcal{E}\{y_m\}$$

Application to uncertainty

$$\mathcal{E}\{s^2\} = \mathcal{E}\left\{\sum_m \sum_{m'} b_{mm'} y_m y_{m'}\right\} = \sum_i \sum_j b_{mm'} \mathcal{E}\{y_m y_{m'}\}$$

can be attributed to pairs of features

Bley et al. Pattern Recognition 2025

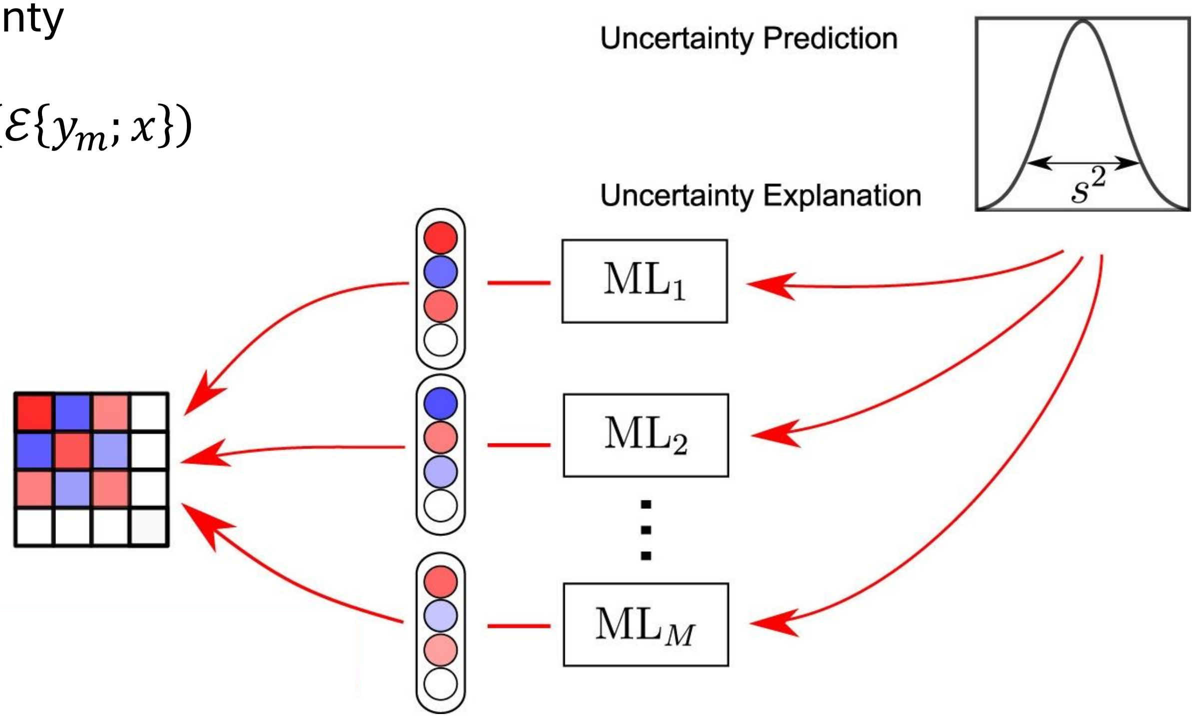
Explaining Uncertainty

Explanation of products

$$\mathcal{E}\{y_m y_{m'}; xx^\top\} = \mathcal{E}\{y_m; x\} \otimes \mathcal{E}\{y_{m'}; x\}$$

Application to uncertainty

$$\mathcal{E}\{s^2; xx^\top\} = \mathrm{Cov}_m(\mathcal{E}\{y_m; x\})$$



Bley et al. Pattern Recognition 2025

Evaluating Explanation Fidelity

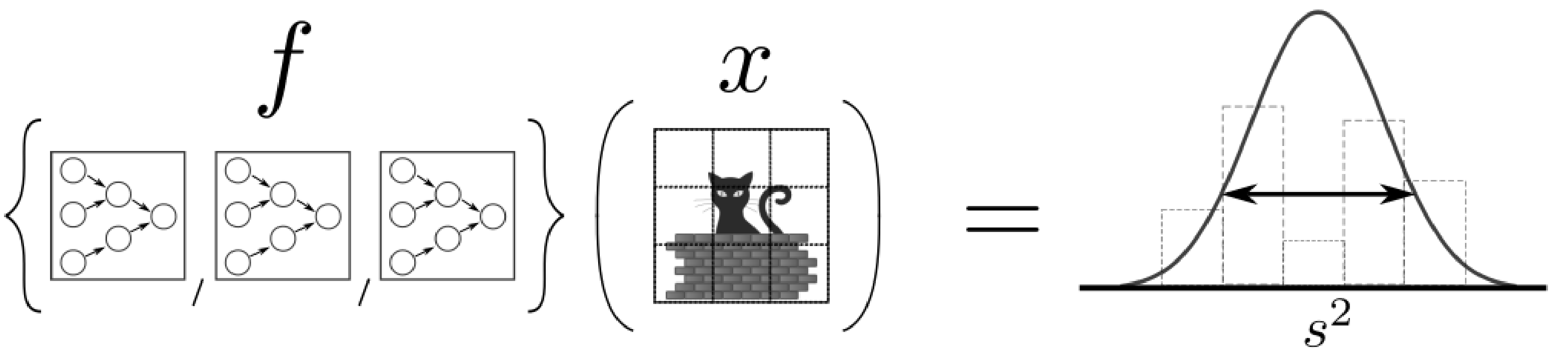
$$\mathcal{E}\{s^2\} = \mathrm{Var}_m(\mathcal{E}\{y_m(x)\}) \qquad \mathcal{E}\{s^2\} = \mathrm{Cov}_m(\mathcal{E}\{y_m(x)\})$$

Dataset (d)	Model	CovLRP		LRP	GI	IG	SVS
		diag	marg				
Bias Correction (21)	DeepEns	**0.352**	0.444	0.411	0.559	0.546	0.513
California Housing (8)	DeepEns	**0.344**	0.370	0.415	0.430	0.394	0.391
EPEX-FR (96)	DeepEns	**0.044**	0.052	0.106	0.113	0.099	0.062
kin8nm (8)	DeepEns	0.391	0.388	0.462	0.427	0.405	0.386
Seoul Bike Sharing (98)	DeepEns	**0.268**	0.294	0.293	0.350	0.338	0.329
Wine Quality (11)	DeepEns	0.482	**0.471**	0.526	0.517	0.500	0.495
YearPredictionMSD (90)	DeepEns	**0.155**	0.173	0.184	0.264	0.273	0.195
Bias Correction	MCDropout	**0.514**	0.517	0.568	0.651	0.530	0.672
California Housing	MCDropout	**0.674**	0.691	0.728	0.812	0.703	0.787
EPEX-FR	MCDropout	**0.085**	0.091	0.137	0.146	0.119	0.125
kin8nm	MCDropout	**0.483**	0.486	0.568	0.586	0.498	0.593
Seoul Bike Sharing	MCDropout	**0.520**	0.590	0.555	0.640	0.568	0.676
Wine Quality	MCDropout	0.661	**0.657**	0.713	0.729	0.662	0.767
YearPredictionMSD	MCDropout	**0.215**	0.258	0.253	0.391	0.273	0.403
YearPredictionMSD	DeepEns-5	**0.128**	0.148	0.155	0.197	0.212	0.153
YearPredictionMSD	DeepEns-10	**0.155**	0.173	0.184	0.264	0.273	0.195
YearPredictionMSD	DeepEns-20	**0.162**	0.183	0.247	0.250	0.267	0.218
YearPredictionMSD	DeepEns-40	0.180	**0.179**	0.235	0.267	0.277	0.213
EPEX-FR	ConvNet	**0.085**	0.101	0.210	0.159	0.108	0.087
Seoul Bike Sharing	ConvNet	**0.231**	0.308	0.422	0.331	0.306	0.321

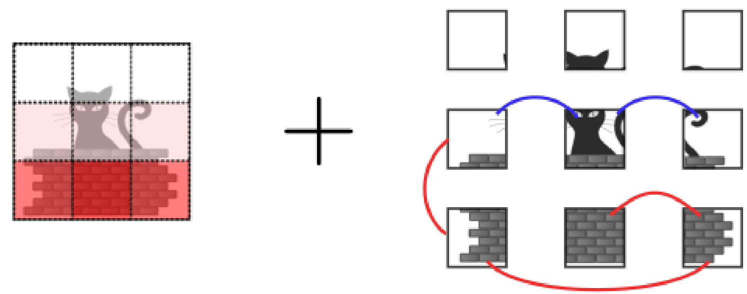
Dropping interaction terms improves explanation robustness.

Bley et al. Pattern Recognition 2025

Explaining Uncertainty: Recap

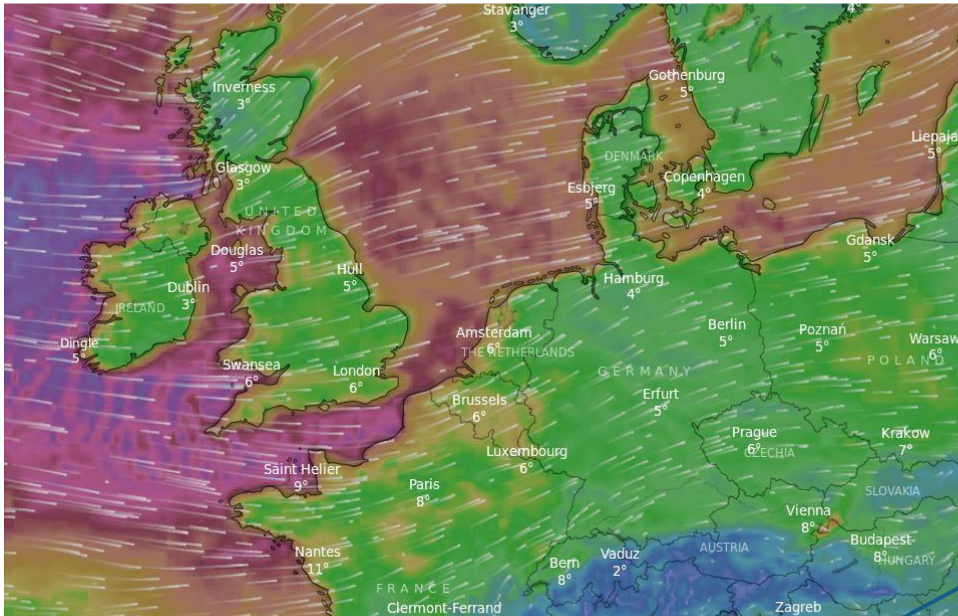


$$\mathcal{E}(s^2; xx^\top) =$$

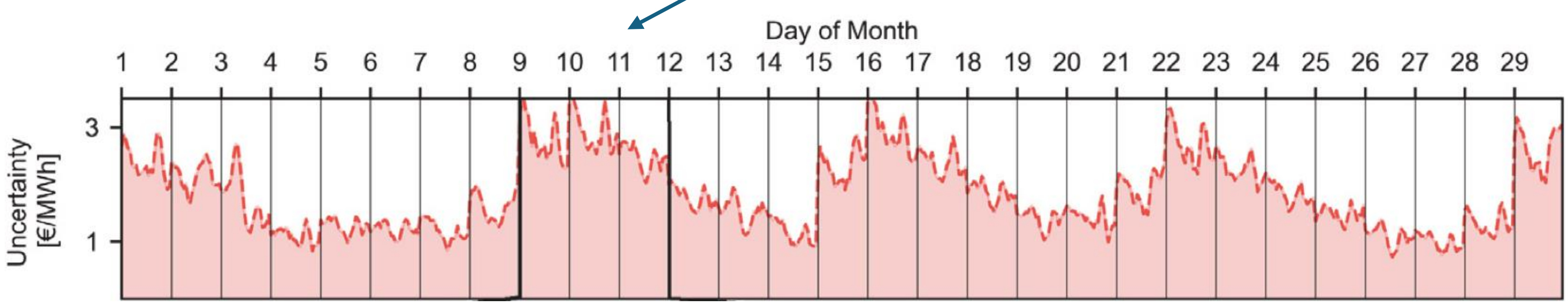


Bley et al. Pattern Recognition 2025

Storm Sabine

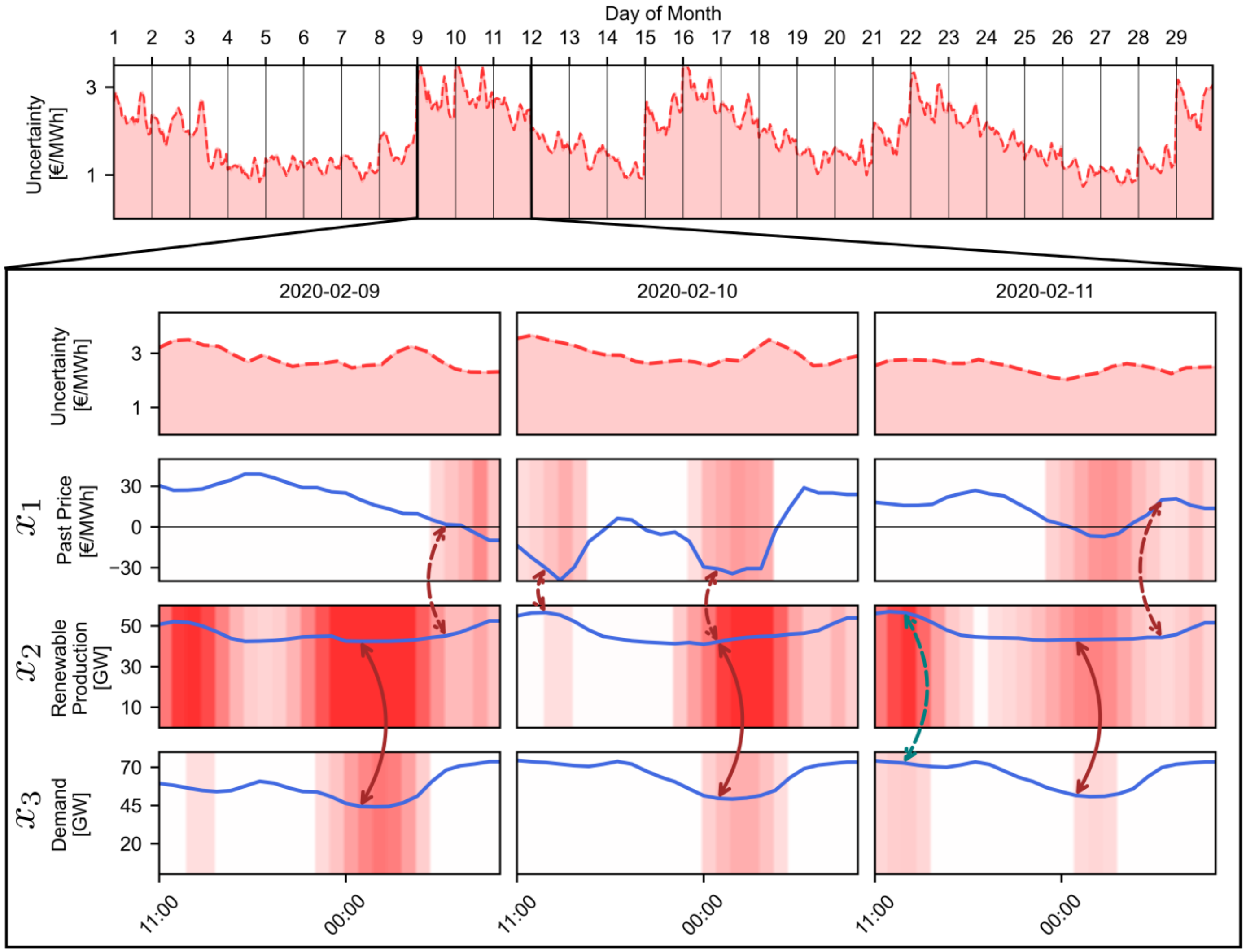


Storm Sabine
9-11 Feb 2020



Bley et al. Pattern Recognition 2025

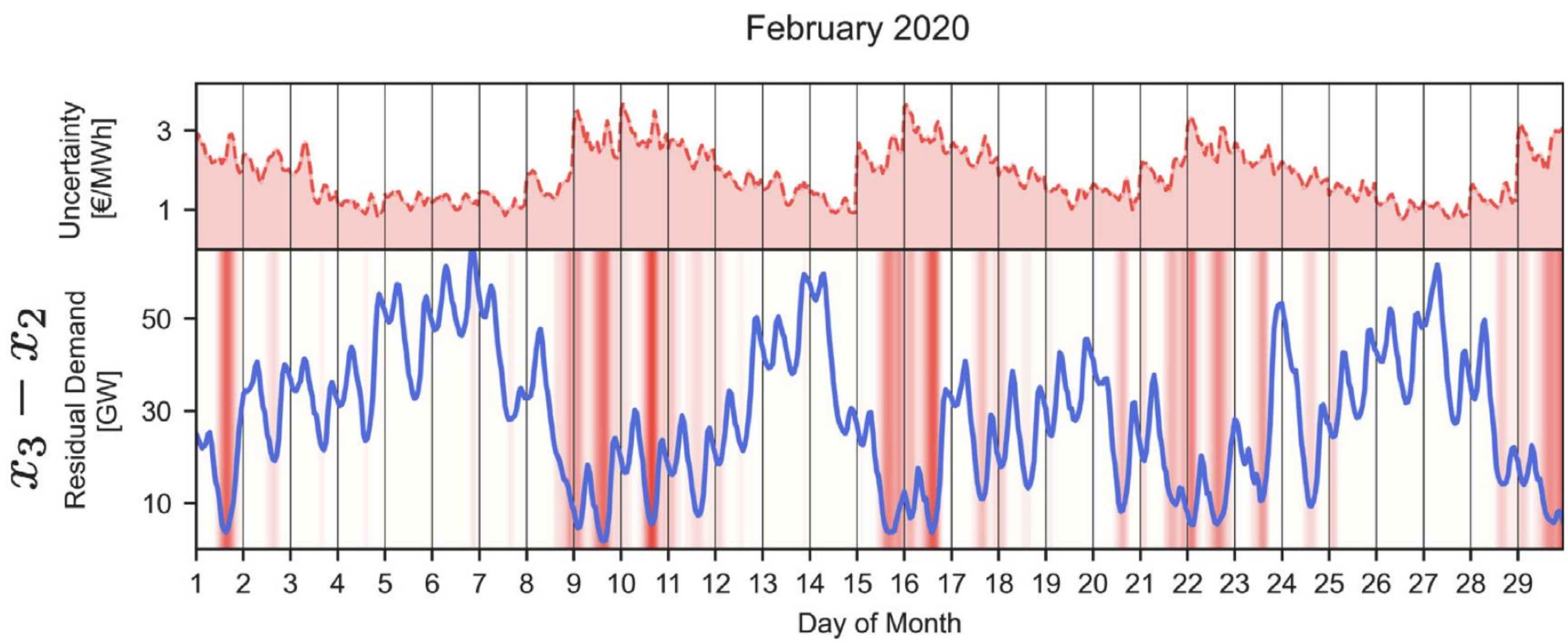
Storm Sabine



$x_3 - x_2$ is the "residual demand"

Bley et al. Pattern Recognition 2025

Storm Sabine



- low residual demand is a clear driver of price uncertainty.
- price uncertainty might further increase due to the growing share of renewables.

Bley et al. Pattern Recognition 2025

Summary

Summary

- With a lot of data, powerful ML models, and with the additional help of deconfounding techniques, many confounding effects can be avoided.
- Recent deep neural networks provide evidence for increased focus on causal features, making observational studies increasingly attractive.
- XAI can adapt to a wide range of ML models and tasks beyond classification (e.g. explaining uncertainty predictions).

Thanks

- P Chormai, J Herrmann, KR Müller, G Montavon. Disentangled explanations of neural network predictions by finding relevant subspaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 46 (11), 7283-7299, 2024

- S Bender, O Delzer, J Herrmann, HA Marxfeld, KR Müller, G Montavon. Mitigating Clever Hans Strategies in Image Classifiers through Generating Counterexamples. arXiv:2510.17524, 2025

- J Keyl, P Keyl, G Montavon, R Hosch, A Brehmer, L Mochmann, ... Decoding pan-cancer treatment outcomes using multimodal real-world data and explainable artificial intelligence. Nature Cancer 6 (2), 307-322, 2025

- 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

- F Bley, S Lapuschkin, W Samek, G Montavon. Explaining predictive uncertainty by exposing second-order effects. Pattern Recognition 160, 111171, 2024