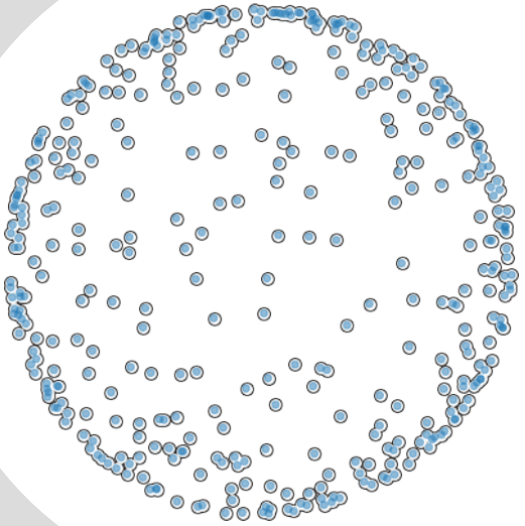
WiSe 2024/25

**Machine Learning for Data Science**

Lecture by G. Montavon

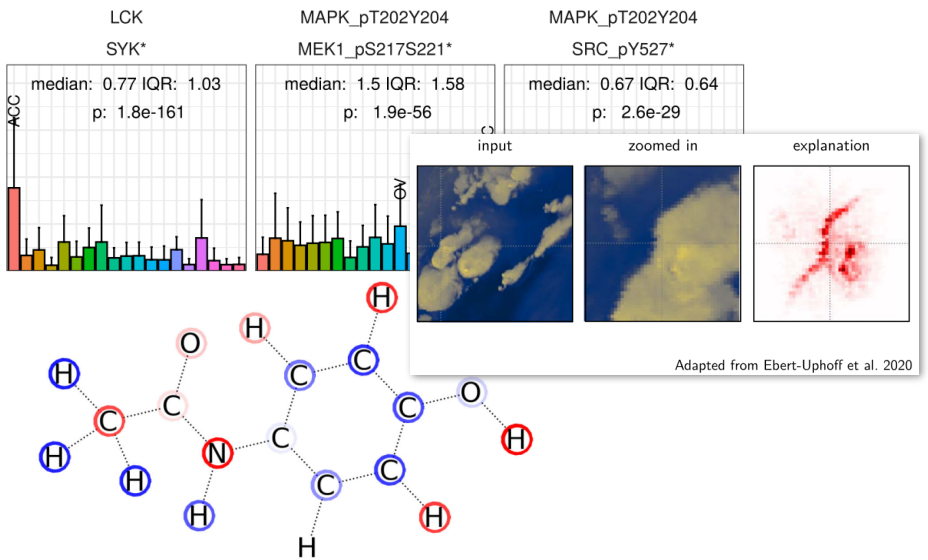
FREIE UNIVERSITÄT BERLIN



Lecture 12b

# Explainable AI

# Beyond 'Classical' Explainable AI

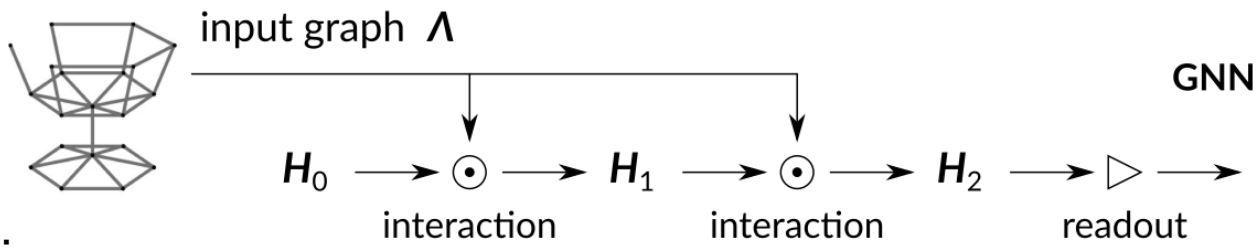



- Current explainable AI already provides single-instance nonlinear explanation capabilities that exceed by far classical statistical measures such as correlation.
- There is a potential demand for even more detailed explanations (e.g. *joint* features contributions, or latent concepts underlying features contributions).

Part 1

# Higher-Order Explanations

# XAI for Graphs (Schnake et al. 2022)




**Observation:**

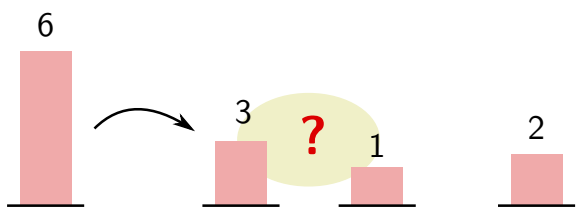
- Input of a GNN is not at layer one, but occurs (multiplicatively) at each layer.

# Limits of 'Classical' Attributions

Function evaluation:

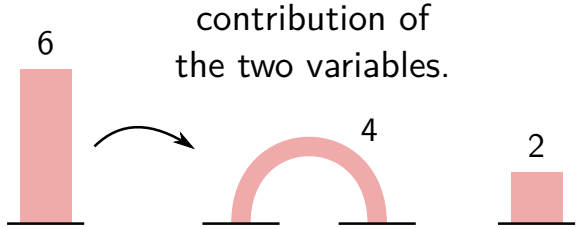
$$\overbrace{f(\boldsymbol{x})}^{6} = \overbrace{x_1}^{4} \cdot \overbrace{x_2}^{1} + \overbrace{x_3}^{2}$$

'Classical' attribution:




better modeled as the joint contribution of the two variables.

Higher-order attribution:

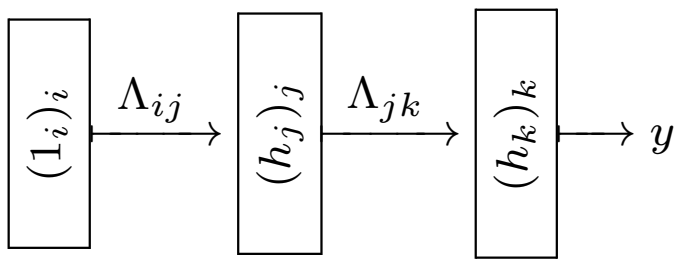



Choice between first-order and higher-order is determined by the *model* rather than by the user.

# XAI for Graphs (Schnake et al. 2022)

GNN prediction (simplified):

$$h_j = \rho\big(\textstyle\sum_i 1_i \Lambda_{ij} w_j\big) \qquad \text{(layer 1)}$$
$$h_k = \rho\big(\textstyle\sum_j h_j \Lambda_{jk} w_k\big) \qquad \text{(layer 2)}$$
$$y = \textstyle\sum_k h_k \qquad \text{(layer 3)}$$



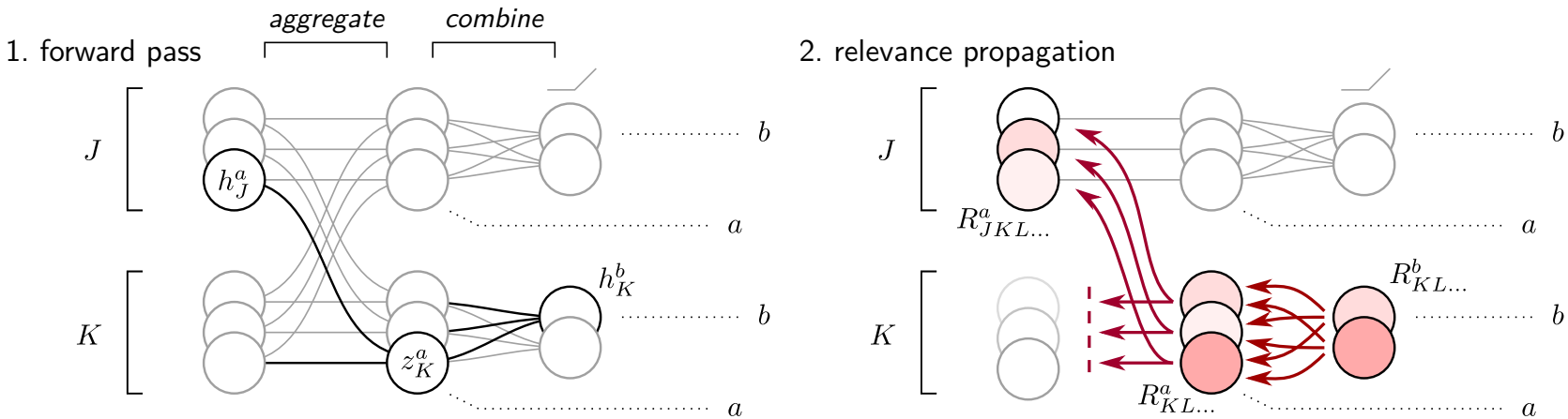

**Our approach:** computing $R_{ijk}$ iteratively:

$$R_{jk} = \mathcal{E}(y, \Lambda_{jk}) \qquad \text{(step 1)}$$
$$R_{ijk} = \mathcal{E}(R_{jk}, \Lambda_{ij}) \qquad \text{(step 2)}$$

**Property:** For $\rho$ linear, the iterative attribution produces the same result as identifying the summands in the expanded form:

$$y = \sum_{ijk} \underbrace{\Lambda_{ij}\Lambda_{jk} 1_i w_j w_k}_{R_{ijk}}$$

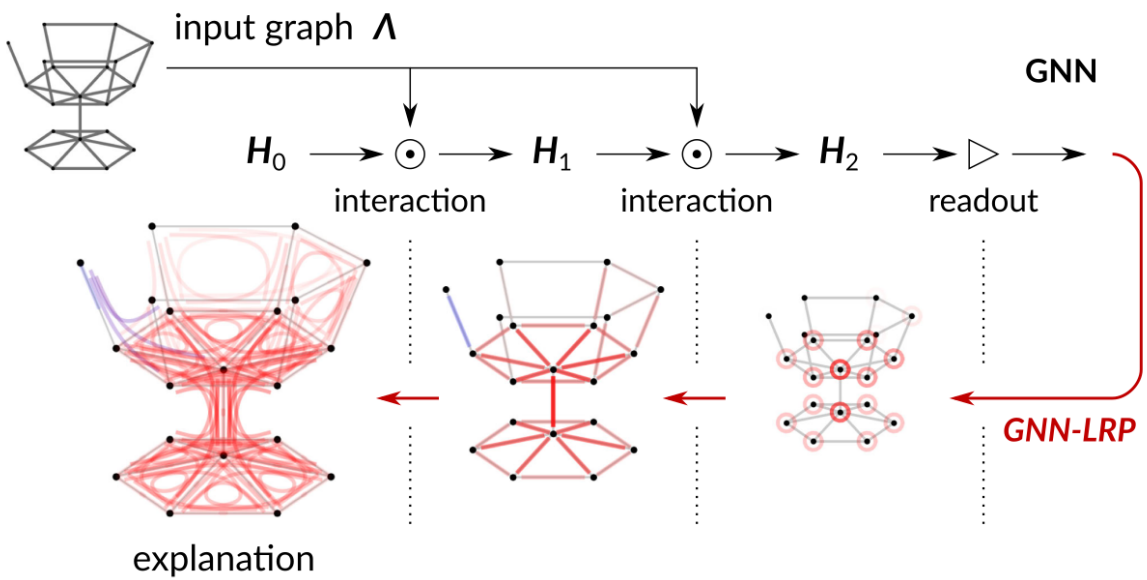
# XAI for Graphs (Schnake et al. 2022)




| Model | Aggregate | Combine | GNN-LRP Rule |
|---|---|---|---|
| GCN [36] | $\boldsymbol{Z}_t = \boldsymbol{\Lambda}\boldsymbol{H}_{t-1}$ | $\boldsymbol{H}_t = \rho(\boldsymbol{Z}_t\boldsymbol{W}_t)$ | $R_{JKL...}^a = \sum_b \frac{\lambda_{JK} h_J^a w_{ab}^{\uparrow}}{\sum_{J,a} \lambda_{JK} h_J^a w_{ab}^{\uparrow}} R_{KL...}^b$ (11) |
| GIN [44] | $\boldsymbol{Z}_t = \boldsymbol{\Lambda}\boldsymbol{H}_{t-1}$ | $\boldsymbol{H}_t = (\mathrm{MLP}^{(t)}(\boldsymbol{Z}_{t,K}))_K$ | $R_{JKL...}^a = \sum_b \frac{\lambda_{JK} h_J^a}{\sum_J \lambda_{JK} h_J^a} \mathrm{LRP}(R_{KL...}^b, z_K^a)$ (12) |
| Spectral [43], [45] (case $\lambda \geq 0$) | $\boldsymbol{Z}_{s,t} = \boldsymbol{\Lambda}_s \boldsymbol{H}_{t-1}$ | $\boldsymbol{H}_t = \rho(\sum_s \boldsymbol{Z}_{s,t}\boldsymbol{W}_{s,t})$ | $R_{JKL...}^a = \sum_b \frac{\sum_s \lambda_{JK}^s h_J^a w_{ab}^{s\uparrow}}{\sum_{J,a} \sum_s \lambda_{JK}^s h_J^a w_{ab}^{s\uparrow}} R_{KL...}^b$ (13) |

# XAI for Graphs (Schnake et al. 2022)

GNN-LRP at work:



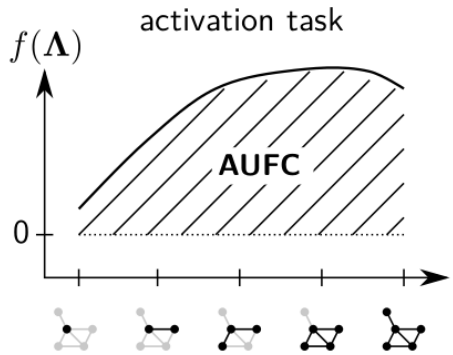

Note:

- In vanilla form, GNN-LRP requires an LRP pass for each walk in the graph ($\rightarrow$ expensive).
- Coarse-graining of the input graph can reduce computations.

# Evaluating Higher-Order Explanations (Schnake et al. 2022)

**Observation:**

- XAI evaluation techniques such as 'Pixel-Flipping' require as input a sequence of features (e.g. nodes) from most to least relevant. However, Higher-Order XAI attributes to joint features.




**Idea:**

- From the given explanation, generalize relevance scores to subset of features $\mathcal{S}$:

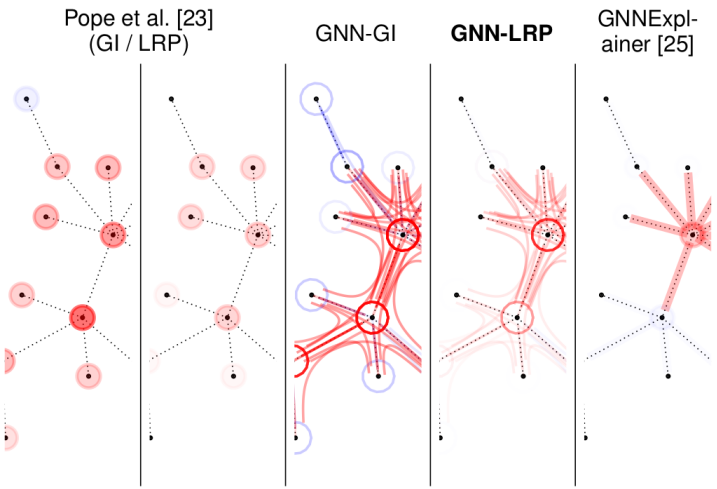
$$R_{\mathcal{S}} = \sum_{i \in \mathcal{S}} R_i \quad \text{(first-order XAI)} \qquad R_{\mathcal{S}} = \sum_{(ijk) \subseteq \mathcal{S}} R_{ijk} \quad \text{(higher-order XAI)}$$

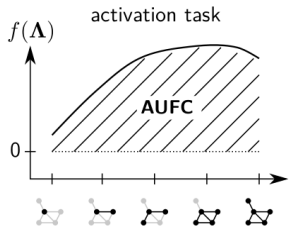
- Ask the explanation to produce an optimal sequence of nodes:

$$\mathcal{Q} = \underset{\mathcal{S}_1 \subset \cdots \subset \mathcal{S}_d}{\operatorname{argmax}} \left\{ \sum_{i=1}^{d} R_{\mathcal{S}_i} \right\}$$

- Finding $\mathcal{Q}$ is intractable $\Rightarrow$ approximate it with greedy feature selection or randomization.

# Evaluating Higher-Order Explanations (Schnake et al. 2022)



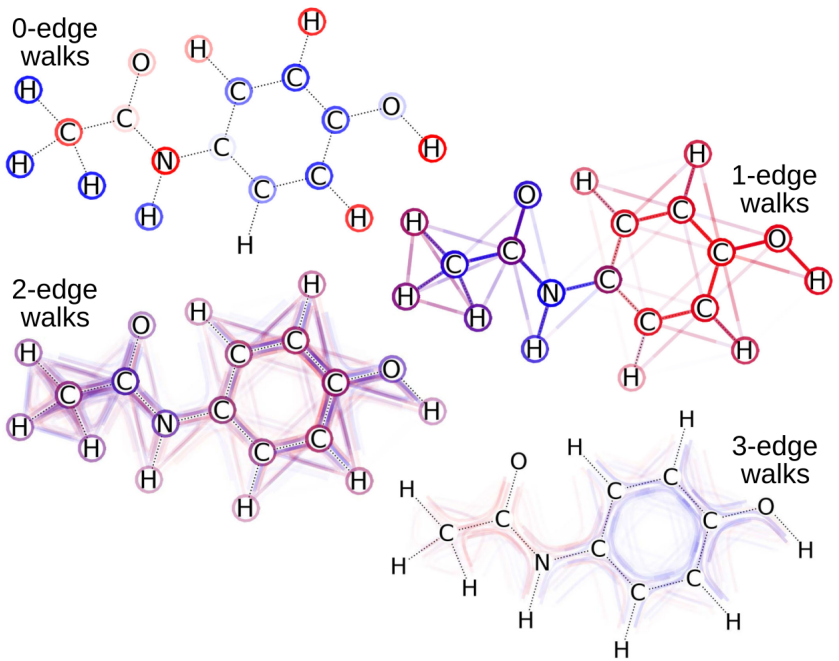

| | P [24] (GI) | P [24] (LRP) | GNN-GI (ours) | GNN-LRP (ours) | GNNExpl [26] | Random |
|---|---|---|---|---|---|---|
| BA-growth, GCN | 2.54 | 3.02 | 2.93 | **3.52** | 3.32 | 1.05 |
| BA-growth, GIN | 2.75 | 3.49 | 3.04 | **3.84** | 3.71 | 1.18 |
| BA-growth, spectral | 0.29 | **1.91** | 0.23 | 1.85 | 1.65 | 0.09 |
| SST, GCN | 26.07 | 26.35 | 26.40 | 26.65 | **27.07** | 20.47 |
| SchNet-$E$ [37] | 10.39 | | **10.47** | | 10.41 | 8.00 |
| SchNet-$\mu$ [37] | 0.87 | | **1.09** | | 1.01 | 0.38 |
| VGG-16 [49] | 9.46 | 13.18 | 12.03 | **14.04** | — | 7.90 |

**Results:**

- GNN-LRP achieves better performance than first-order explanations (LRP and GNNExpl).
- GNN-LRP is more robust than its simpler gradient-based counter part GNN-GI.

# Use Case: XAI for Quantum Chemistry

Decomposing molecular properties (predicted via a GNN) in terms of atom interactions of different order.



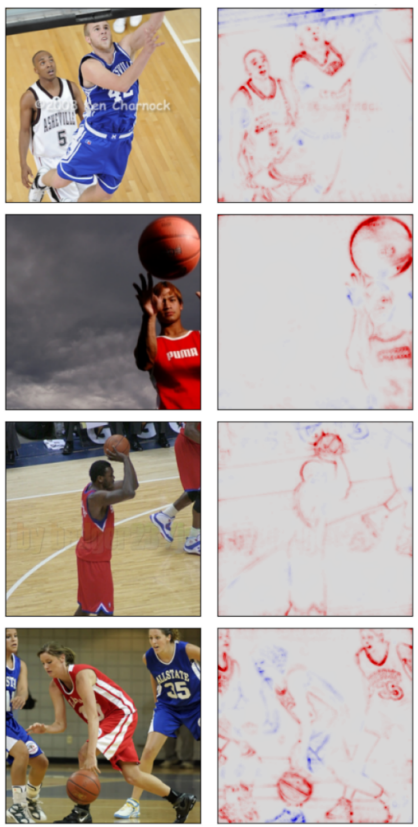

**Challenges:**

- Larger explanations → more difficult to comprehend for a human.
- General comment about XAI: Need to make a distinction between the strategy employed by the model to predict (dataset-specific) and the underlying physics (general).

Part 2

# Disentangled Explanations

# Limits of 'Classical' Explanations

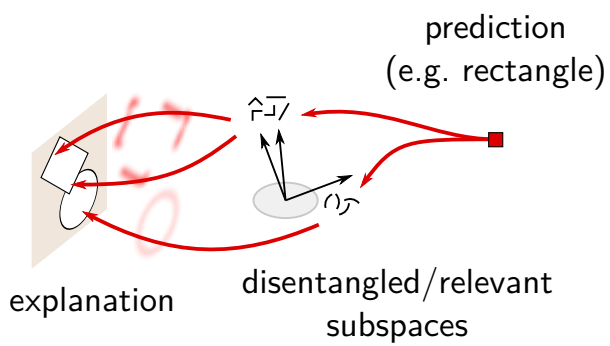


**Observation:**

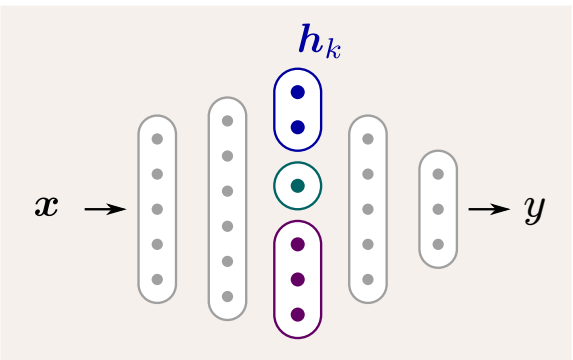
- Several concepts (ball, player, etc.) are entangled in the same explanation.

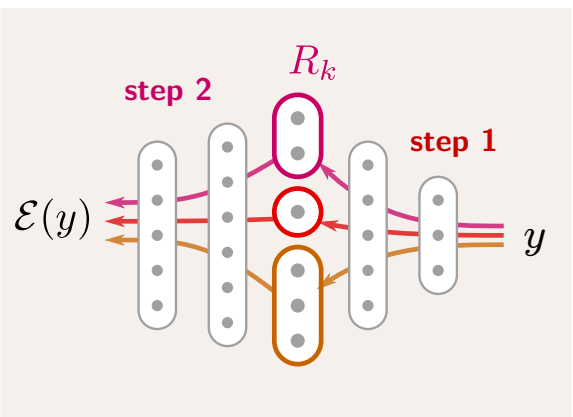
**Question:**

- Can we disentangle explanations into multiple distinct concepts so that they become more actionable?



# Disentangled Explanations (Chormai et al. 2024)




Forward pass:

$$\boldsymbol{x} \mapsto (h_k)_k \qquad \text{(input to subspaces)}$$
$$(h_k)_k \mapsto y \qquad \text{(subspaces to output)}$$

Standard explanation

$$R_i = \mathcal{E}(y, x_i)$$

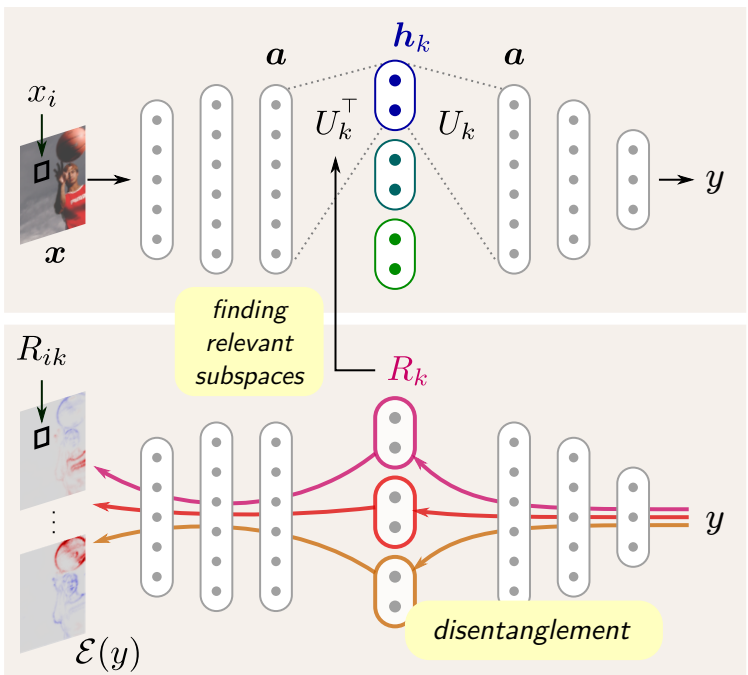
Disentangled explanation (ours):

$$R_k = \mathcal{E}(y, h_k) \qquad \text{(step 1)}$$
$$R_{ik} = \mathcal{E}(R_k, x_i) \qquad \text{(step 2)}$$

# Extracting Relevant Subspaces (Chormai et al. 2024)

**Notation:**

| | |
|---|---|
| $\boldsymbol{a}$ | Vector of activations |
| $\boldsymbol{R}$ | Vector of activation relevances |
| $\boldsymbol{c}$ | Vector such that $\boldsymbol{R} = \boldsymbol{a} \odot \boldsymbol{c}$ |
| $(U_k)_k$ | Matrices that project activations to orthogonal subspaces.) |




**Key findings:**

1. For a variety of methods (e.g. integrated gradients, LRP), the relevance score for subspace $k$ can be expressed as:

$$R_k = (U_k^\top \boldsymbol{a})^\top (U_k^\top \boldsymbol{c})$$

2. We can find subspaces that *directly* maximize some statistic of $R_k$.

# Two Proposed Analyses (Chormai et al. 2024)

## Principal Relevant Component Analysis (PRCA)

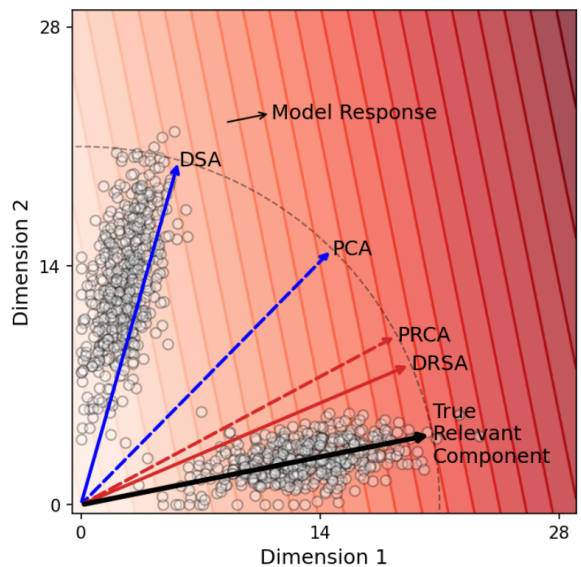
$$\underset{U}{\text{maximize}}: \overbrace{\text{Tr}(U^\top \underbrace{\mathbb{E}[\boldsymbol{a}\boldsymbol{c}^\top]}_{\Sigma_{\boldsymbol{a}\boldsymbol{c}}} U)}^{R}$$

If setting $\boldsymbol{c} \leftarrow \boldsymbol{a}$, PRCA reduces to (uncentered) PCA.

## Disentangled Relevant Subspace Analysis (DRSA)

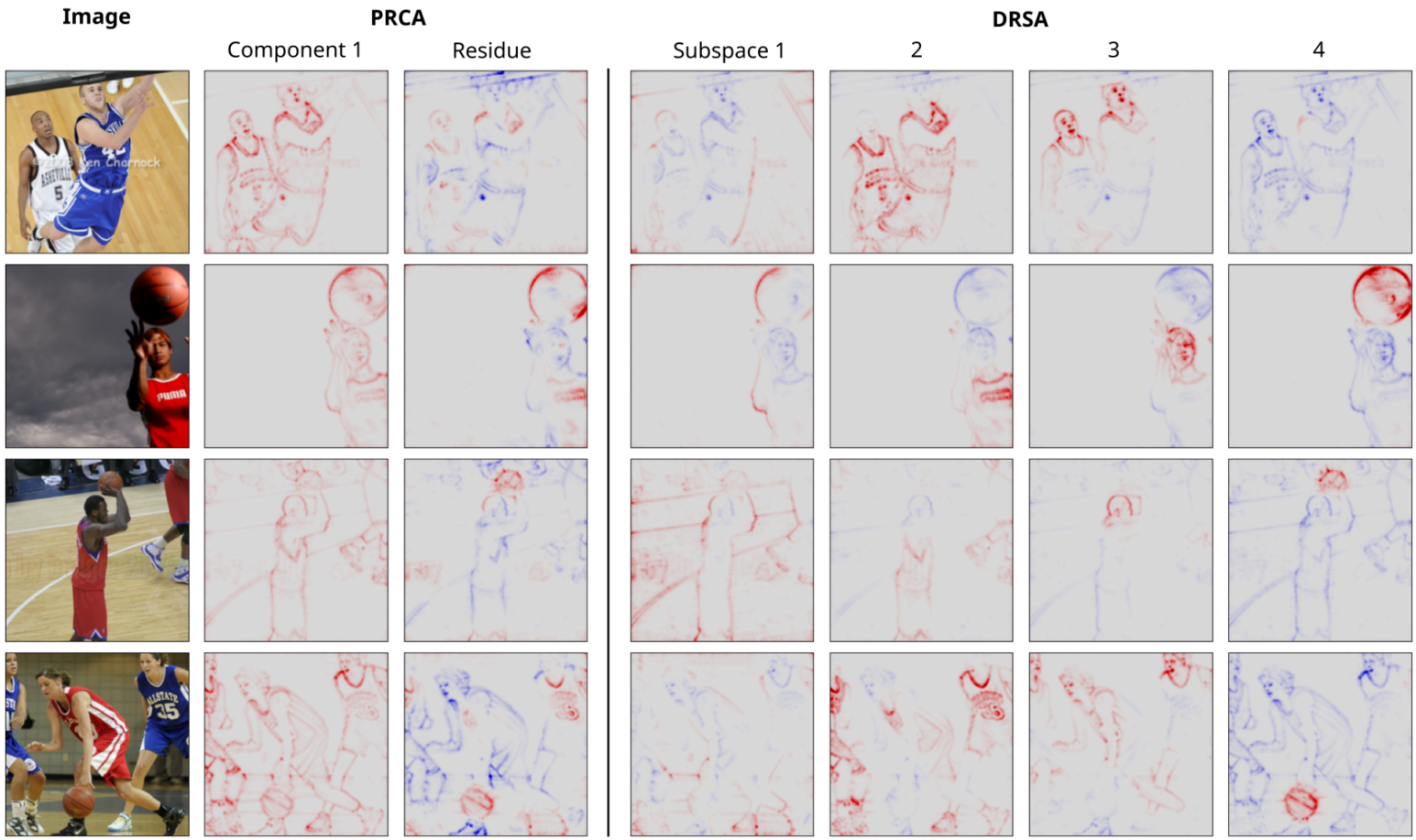
$$\underset{(U_k)_k}{\text{maximize}}: \mathbb{M}_k^{0.5} \mathbb{M}_n^2 \{ ( \overbrace{(U_k^\top \boldsymbol{a}_n)^\top (U_k^\top \boldsymbol{c}_n)}^{R_{kn}} )^+ \}$$

If setting $\boldsymbol{c} \leftarrow \boldsymbol{a}$ DRSA reduces to 'DSA'.




Unlike PCA/ICA/DSA/..., our analyses focus on components that are *relevant* for the prediction.

# PRCA/DRSA in Practice (Chormai et al. 2024)



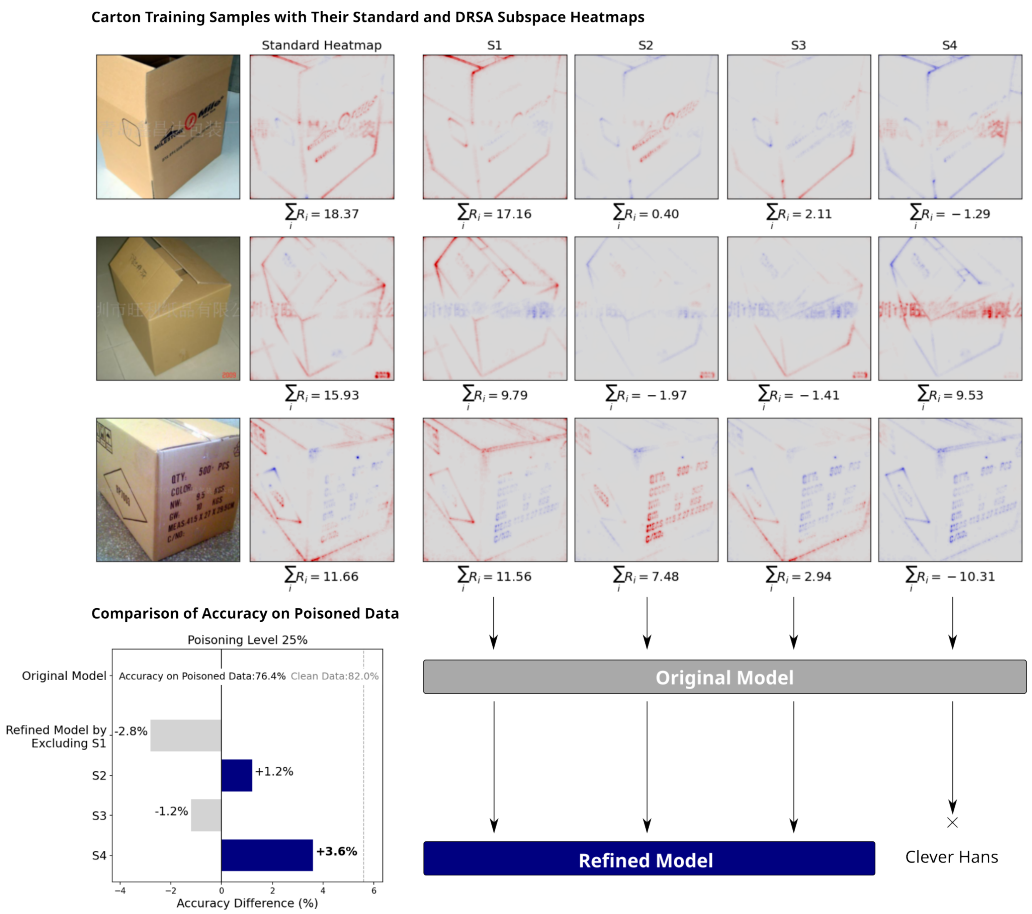
# Use Case: Detecting and Removing Clever Hanses

**Current approaches:**

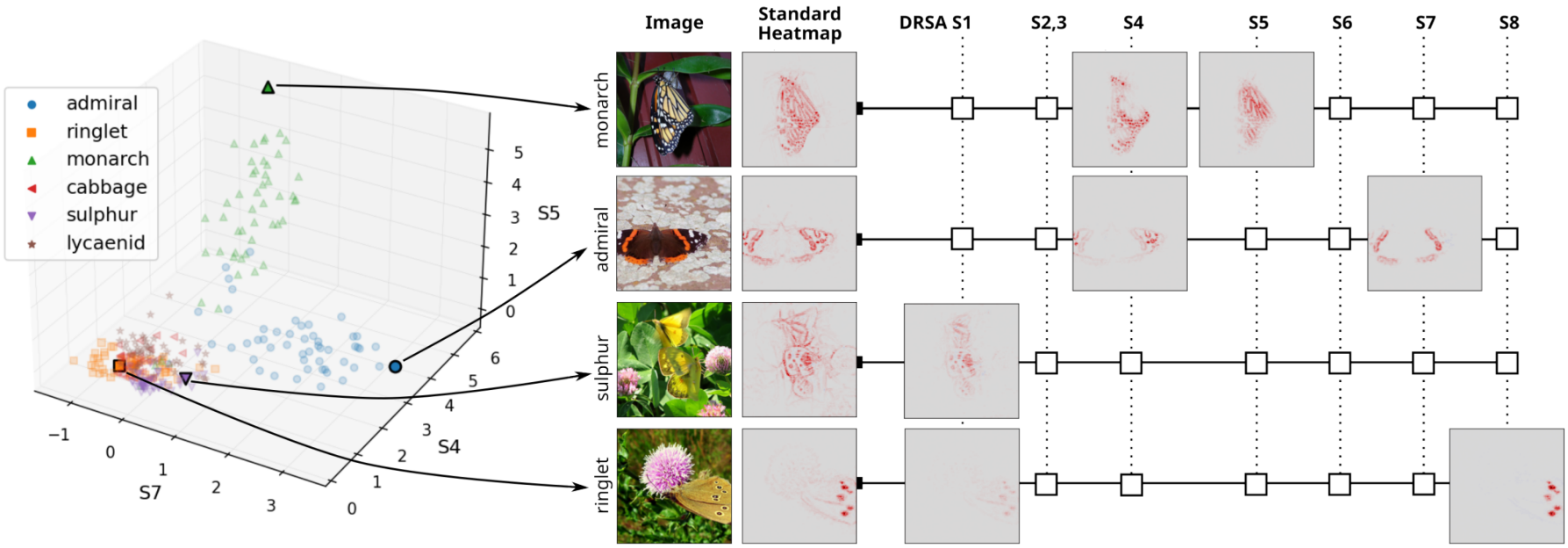
- Artifact models built from preliminarily detected Clever Hans instances.

**Our approach:**

1. Observe that Clever Hans strategies readily occur in distincts components of DRSA.
2. Identify these components, and remove their contribution from the overall prediction.



# Use Case: Exploring Visual Relations between Classes




- Certain visual concepts are shared between classes (e.g. dotted pattern of 'admiral' and 'monarch' butterflies).
- This can be analyzed dataset-wide in a scatter plot (left).

# Summary

# Summary

- Explanations should not only be faithful/understandable; they should also be informative & actionable by the user.

- Possible directions to achieve this:
  - Ensuring the explanation reflects the use of **higher-order** feature interactions by the model (e.g. GNN-LRP).
  - Resolving the latent concepts attached to each feature contribution in order to produce a **disentangled** explanation (e.g. using PRCA / DRSA).

# References


P. Chormai, J. Herrmann, K.-R. Müller, and G. Montavon.

Disentangled explanations of neural network predictions by finding relevant subspaces.

*IEEE Trans. Pattern Anal. Mach. Intell.*, 46(11):7283–7299, 2024.

T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K. T. Schütt, K. Müller, and G. Montavon.

Higher-order explanations of graph neural networks via relevant walks.

*IEEE Trans. Pattern Anal. Mach. Intell.*, 44(11):7581–7596, 2022.