IPAM Workshop
Explainable AI for the Sciences: Towards Novel Insights

Towards Higher-Order & Disentangled XAI

Grégoire Montavon
gregoire.montavon@fu-berlin.de

11 January 2023

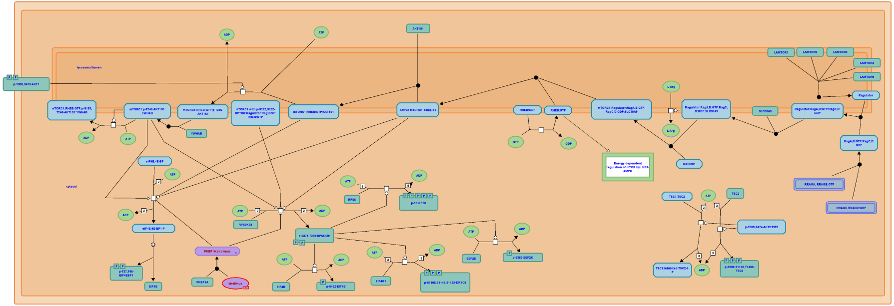
Freie Universität Berlin BIFOLD

Part 1: Motivations

P Keyl, M Bockmayr, D Heim, G Dernbach, G Montavon, KR Müller, F Klauschen
Patient-level proteomic network prediction by explainable artificial intelligence
NPJ Precis Oncol. 6(1):35, 2022

Example: Discovering Influencial Proteins

Example: MTOR signaling network (from reactome.org)



MTOR signalling

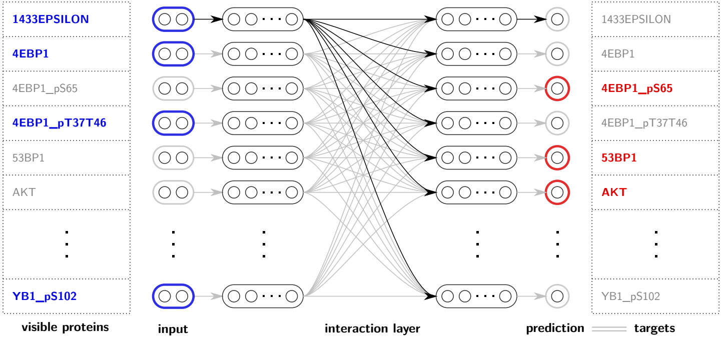
Question:

- Can we use ML/XAI to infer these networks (or aspects of them) directly from the data?

Finding Influential Proteins with ML/XAI (Keyl et al. 2022)

Step 1: From Data to ML

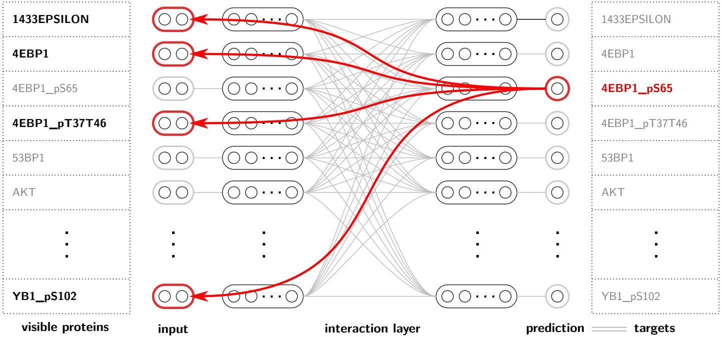
- Assemble a dataset.
- Build a ML model (neural network) that predicts proteins from other proteins with best possible accuracy.



Finding Influential Proteins with ML/XAI (Keyl et al. 2022)

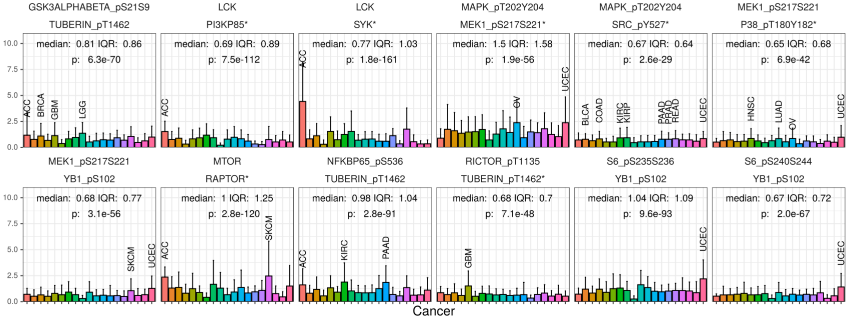
Step 2: From ML to XAI

- Apply Explainable AI (here the LRP attribution technique) to identify to what extent proteins contribute to the expression of other proteins.

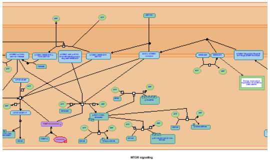


Finding Influential Proteins with ML/XAI (Keyl et al. 2022)

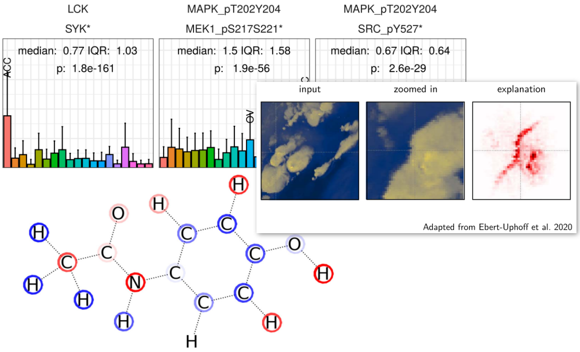
Excerpt of top-k protein influences via XAI (LRP attribution):



- Generally consistent with existing knowledge, e.g. highlights mTOR pathway; correlates with entries in the reactome knowledgebase (`https://reactome.org/`).
- Provides *cancer-specific* (or even *instance-specific*) view of protein influences.



Beyond ‘Classical’ Explainable AI

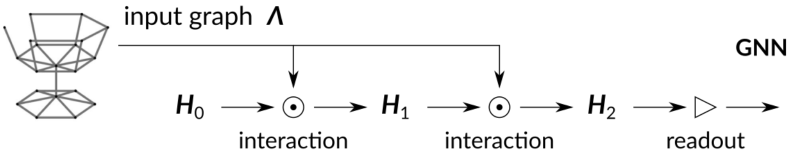


- Current explainable AI already provides single-instance nonlinear explanation capabilities that exceed by far classical statistical measures such as correlation.
- There is a potential demand for even more detailed explanations (e.g. *joint* features contributions, or latent concepts underlying features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

Part 2: Towards Higher-Order Explainable AI

T Schnake, O Eberle, J Lederer, S Nakajima, K T. Schütt, KR Müller, G Montavon
Higher-Order Explanations of Graph Neural Networks via Relevant Walks
IEEE TPAMI 44(11):7581-7596, 2022

XAI for Graphs (Schnake et al. 2022)

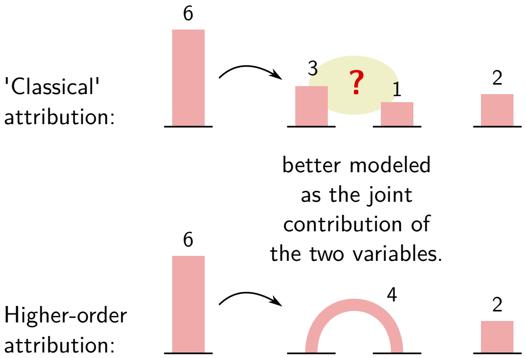


Observation:

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

Limits of ‘Classical’ Attributions

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



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):

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

$$\boxed{(1_i)_i} \xmapsto{\Lambda_{ij}} \boxed{(h_j)_j} \xmapsto{\Lambda_{jk}} \boxed{(h_k)_k} \longmapsto y$$

Our approach: computing R_{ijk} iteratively:

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

XAI for Graphs (Schnake et al. 2022)

GNN prediction (simplified):

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

$$
\boxed{(1_i)_i} \xmapsto{\Lambda_{ij}} \boxed{(h_j)_j} \xmapsto{\Lambda_{jk}} \boxed{(h_k)_k} \longmapsto y
$$

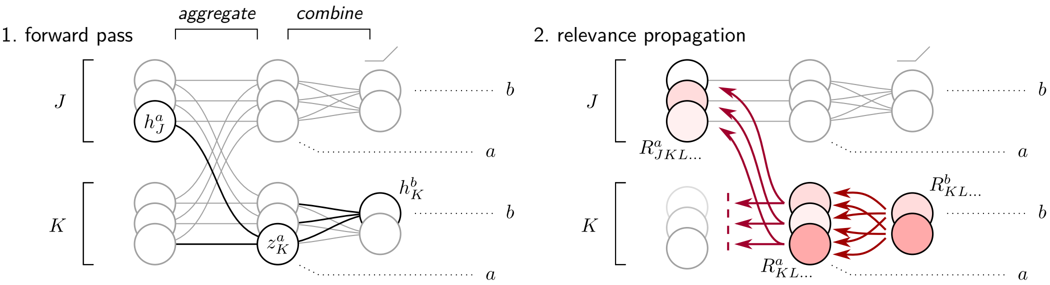
Our approach: computing R_{ijk} iteratively:

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

Property: For ρ linear, the iterative attribution produces the same result as identifying the summands in the expanded form:

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

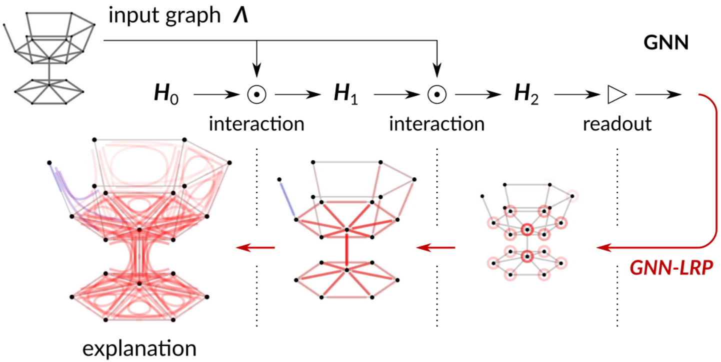
XAI for Graphs (Schnake et al. 2022)



Model	Aggregate	Combine	GNN-LRP Rule	
GCN [36]	$\boldsymbol{Z}_t = \boldsymbol{\Lambda H}_{t-1}$	$\boldsymbol{H}_t = \rho(\boldsymbol{Z}_t \boldsymbol{W}_t)$	$R^a_{JKL\ldots} = \sum_b \frac{\lambda_{JK} h^a_J w^{\uparrow}_{ab}}{\sum_{J,a} \lambda_{JK} h^a_J w^{\uparrow}_{ab}} R^b_{KL\ldots}$	(11)
GIN [44]	$\boldsymbol{Z}_t = \boldsymbol{\Lambda H}_{t-1}$	$\boldsymbol{H}_t = (\mathrm{MLP}^{(t)}(Z_{t,K}))_K$	$R^a_{JKL\ldots} = \sum_b \frac{\lambda_{JK} h^a_J}{\sum_J \lambda_{JK} h^a_J} \mathrm{LRP}(R^b_{KL\ldots}, z^a_K)$	(12)
Spectral [43], [45] (case $\lambda \geq 0$)	$\boldsymbol{Z}_{s,t} = \boldsymbol{\Lambda}_s \boldsymbol{H}_{t-1}$	$\boldsymbol{H}_t = \rho(\sum_s \boldsymbol{Z}_{s,t} \boldsymbol{W}_{s,t})$	$R^a_{JKL\ldots} = \sum_b \frac{\sum_s \lambda^s_{JK} h^a_J w^{s\uparrow}_{ab}}{\sum_{J,a} \sum_s \lambda^s_{JK} h^a_J w^{s\uparrow}_{ab}} R^b_{KL\ldots}$	(13)

XAI for Graphs (Schnake et al. 2022)

GNN-LRP at work:



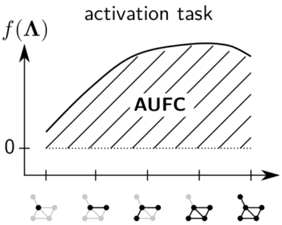
Note:

- 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 $\mathcal{S}$:

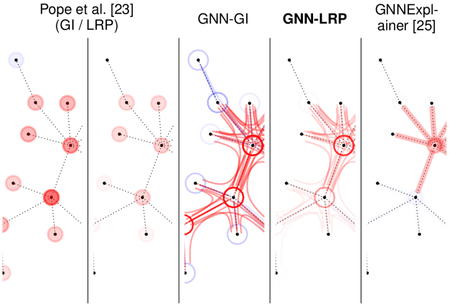
$$R_{\mathcal{S}} = \sum_{i \in \mathcal{S}} R_i \quad \text{(first-order XAI)} \qquad R_{\mathcal{S}} = \sum_{(ijk) \subseteq \mathcal{S}} R_{ijk} \quad \text{(higher-order XAI)}$$

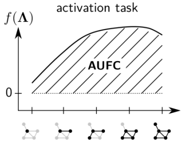
- Ask the explanation to produce an optimal sequence of nodes:

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

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

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





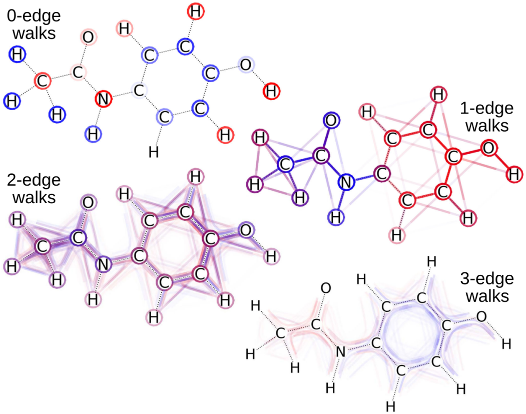
	P [24] (GI)	P [24] (LRP)	GNN-GI (ours)	GNN-LRP (ours)	GNNExpl [26]	Random
BA-growth, GCN	2.54	3.02	2.93	**3.52**	3.32	1.05
BA-growth, GIN	2.75	3.49	3.04	**3.84**	3.71	1.18
BA-growth, spectral	0.29	**1.91**	0.23	1.85	1.65	0.09
SST, GCN	26.07	26.35	26.40	26.65	**27.07**	20.47
SchNet-E [37]	10.39		**10.47**		10.41	8.00
SchNet-μ [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 $\rightarrow$ 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 3: Towards Disentangled Explanations

P Chormai, J Herrmann, KR Müller, G Montavon
Disentangled Explanations of Neural Network Predictions by Finding Relevant Subspaces
arXiv:2212.14855, 2022

Limits of 'Classical' Explanations

Observation:

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

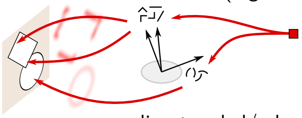
Limits of 'Classical' Explanations

Observation:

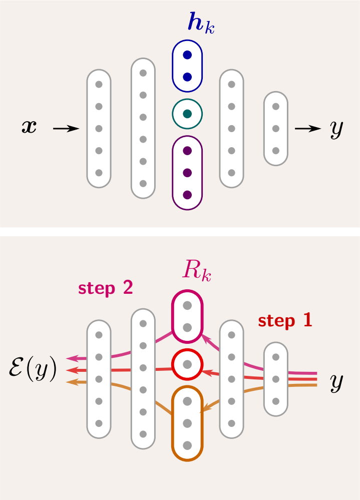
- Several concepts (ball, player, etc.) are entangled in the same explanation.

Question:

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



Disentangled Explanations (Chormai et al. 2022)



Forward pass:

$$\begin{aligned} \boldsymbol{x} &\mapsto (h_k)_k && \text{(input to subspaces)} \\ (h_k)_k &\mapsto y && \text{(subspaces to output)} \end{aligned}$$

Standard explanation

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

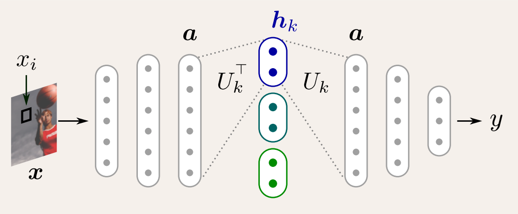
Disentangled explanation (ours):

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

Extracting Relevant Subspaces (Chormai et al. 2022)

Notation:

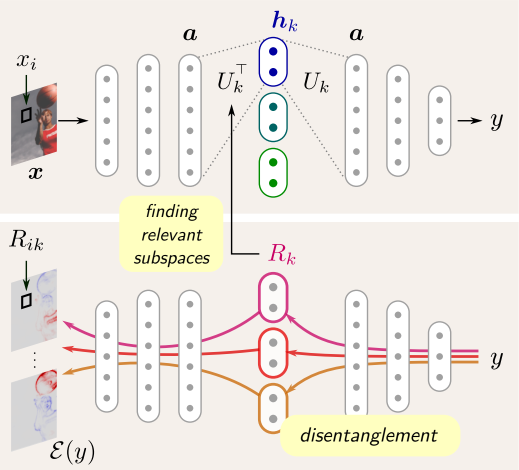
$\boldsymbol{a}$	Vector of activations
$\boldsymbol{R}$	Vector of activation relevances
$\boldsymbol{c}$	Vector such that $\boldsymbol{R} = \boldsymbol{a} \odot \boldsymbol{c}$
$(U_k)_k$	Matrices that project activations to orthogonal subspaces.)



Extracting Relevant Subspaces (Chormai et al. 2022)

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. 2022)

Principal Relevant Component Analysis (PRCA)

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

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

Two Proposed Analyses (Chormai et al. 2022)

Principal Relevant Component Analysis (PRCA)

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

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

Disentangled Relevant Subspace Analysis (DRSA)

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

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

Two Proposed Analyses (Chormai et al. 2022)

Principal Relevant Component Analysis (PRCA)

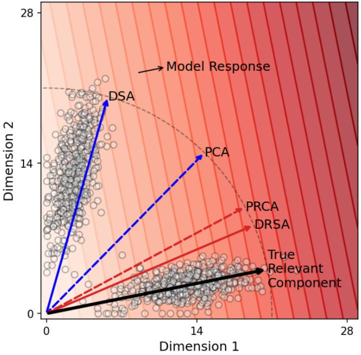
$$\underset{U}{\text{maximize}}: \quad \overbrace{\mathrm{Tr}(U^\top \underbrace{\mathbb{E}[\boldsymbol{a}\boldsymbol{c}^\top]}_{\Sigma_{\boldsymbol{ac}}} U)}^{R}$$

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

Disentangled Relevant Subspace Analysis (DRSA)

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

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



Two Proposed Analyses (Chormai et al. 2022)

Principal Relevant Component Analysis (PRCA)

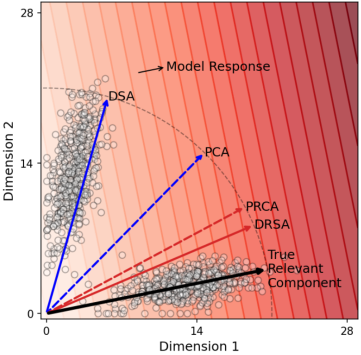
$$\underset{U}{\text{maximize}}: \quad \overbrace{\mathrm{Tr}(U^\top \underbrace{\mathbb{E}[\boldsymbol{a}\boldsymbol{c}^\top]}_{\Sigma_{\boldsymbol{ac}}} U)}^{R}$$

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

Disentangled Relevant Subspace Analysis (DRSA)

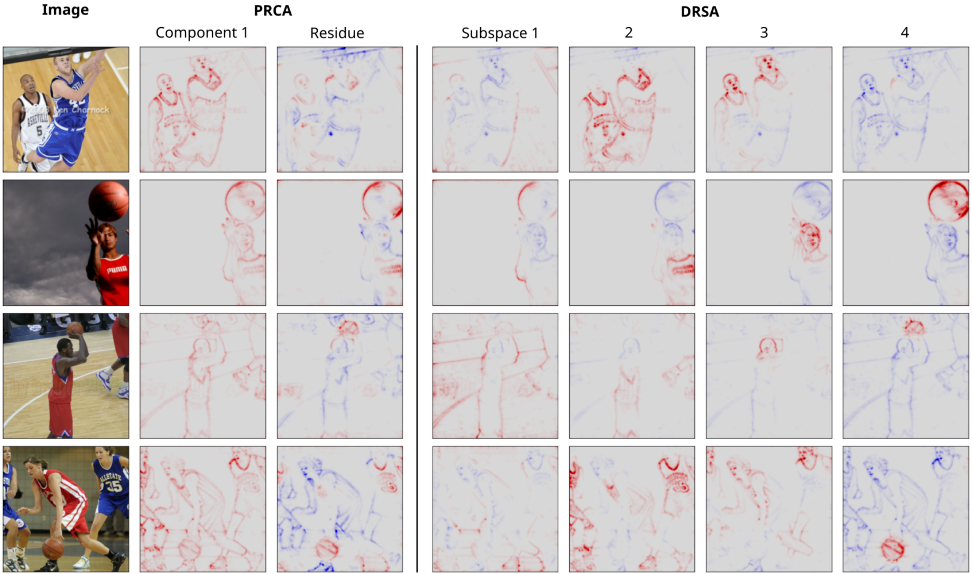
$$\underset{(U_k)_k}{\text{maximize}}: \quad \mathbb{M}_k^{0.5}\mathbb{M}_n^2\Big\{\Big(\overbrace{(U_k^\top \boldsymbol{a}_n)^\top (U_k^\top \boldsymbol{c}_n)}^{R_{kn}}\Big)^+\Big\}$$

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

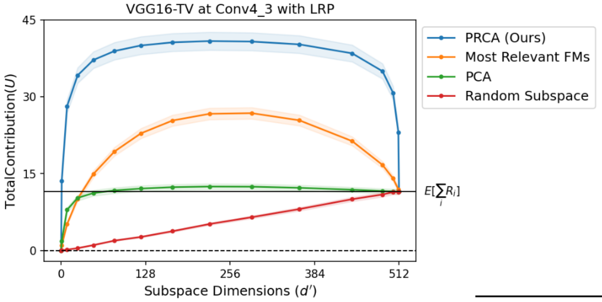


Unlike PCA/ICA/DSA/..., our analyses focus on components that are *relevant* for the prediction.

PRCA/DRSA in Practice (Chormai et al. 2022)



PRCA vs. Baselines (Chormai et al. 2022)

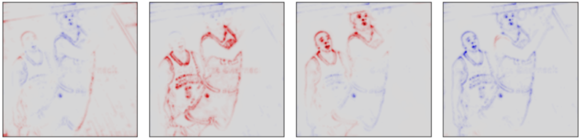


PRCA extracts much more strongly contributing subspaces than baseline methods.

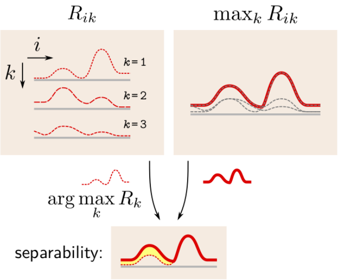
	VGG16-TV + LRP	VGG16-ND + LRP	NFNet-F0 + LRP	VGG16-TV + Shapley	VGG16-ND + Shapley
Total ($\sum_i R_{i,I}$)	*11.47*	*10.35*	*6.57*	*17.23*	*16.59*
Random Subspace	0.02	0.00	0.01	0.02	0.02
Most Relevant FM [22]*	0.97	0.87	0.30	1.12	0.91
PCA	1.81	2.69	-2.22	21.81	18.98
PRCA (Ours)	**13.63**	**13.72**	**11.29**	**44.76**	**42.12**
Error bars (max)	± 0.66	± 0.61	± 0.58	± 1.78	± 1.40

DRSA vs. Baselines (Chormai et al. 2022)

Question: are the components of the explanation spatially disentangled?



Separability score:



	VGG16-TV + LRP	VGG16-ND + LRP	NFNet-F0 + LRP	VGG16-TV + Shapley	VGG16-ND + Shapley
Separability					
Random Subspace	1.00	1.00	1.00	1.00	1.00
NetDissect [46]	1.84	1.94	—	—	—
DSA	2.27	2.25	8.52	1.10	1.05
DRSA (Ours)	**3.28**	**3.03**	**14.40**	**1.78**	**1.63**
Error bars (max)	± 0.17	± 0.15	± 0.89	± 0.07	± 0.07

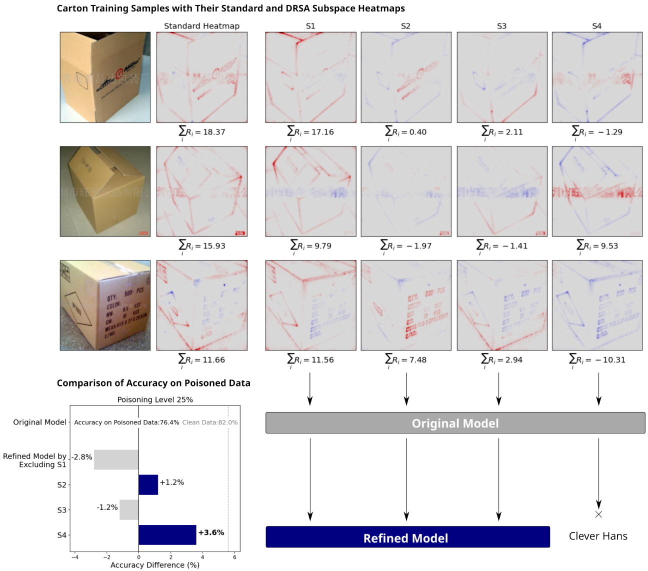
Use Case: Detecting and Removing Clever Hanses

Current approaches:

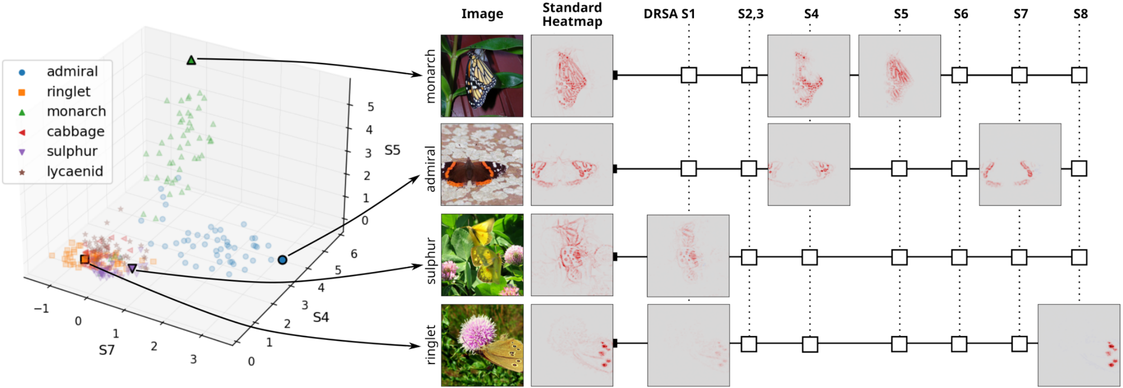
- Artifact models built from preliminarily detected Clever Hans instances.

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

- Explanations should not only be faithful/understandable; they should also be **informative** & **actionable** by the user.
- This can be achieved by:
  - 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).
- Both approaches (higher-order & disentangled XAI) are not mutually exclusive. They could be combined in future work.

References to presented works

XAI for Analyzing Protein Interactions

- P Keyl, M Bockmayr, D Heim, G Dernbach, G Montavon, KR Müller, F Klauschen
  Patient-level proteomic network prediction by explainable artificial intelligence
  NPJ Precis Oncol. 6(1):35, 2022

Higher-Order XAI

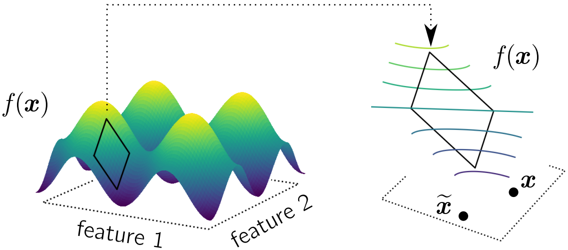
- T Schnake, O Eberle, J Lederer, S Nakajima, K T. Schütt, KR Müller, G Montavon
  Higher-Order Explanations of Graph Neural Networks via Relevant Walks
  IEEE TPAMI 44(11):7581-7596, 2022

Disentangled XAI

- P Chormai, J Herrmann, KR Müller, G Montavon
  Disentangled Explanations of Neural Network Predictions by Finding Relevant Subspaces
  arXiv:2212.14855, 2022

Check our review paper on XAI

- W Samek, G Montavon, S Lapuschkin, C Anders, KR Müller
  Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications
  Proceedings of the IEEE, 109(3):247-278, 2021



Visit our website

- Code/demos for our XAI methods
- Full list of papers