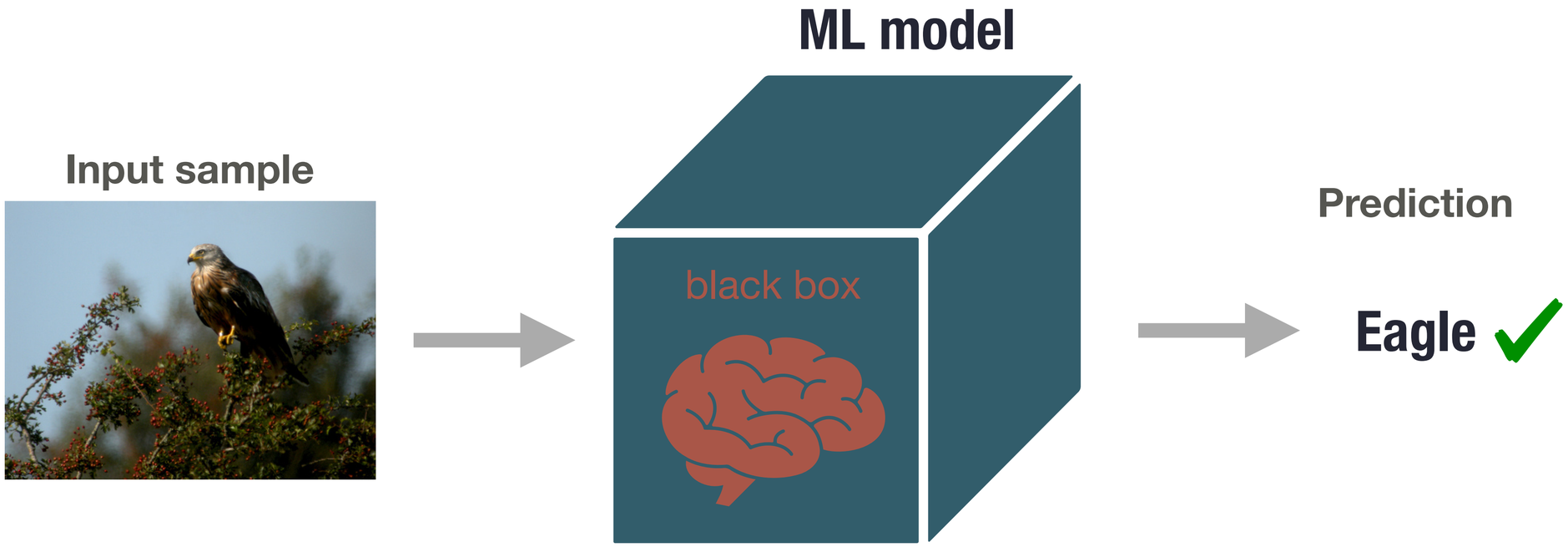
PredDiff: Explanation and Interactions from Conditional Expectations

Stefan Blücher (PhD @ BIFOLD @ Prof. Müller)
Joint work with Johanna Vielhaben and Nils Strodthoff

Warm-up — XAI review

Explainable AI (XAI) — overview



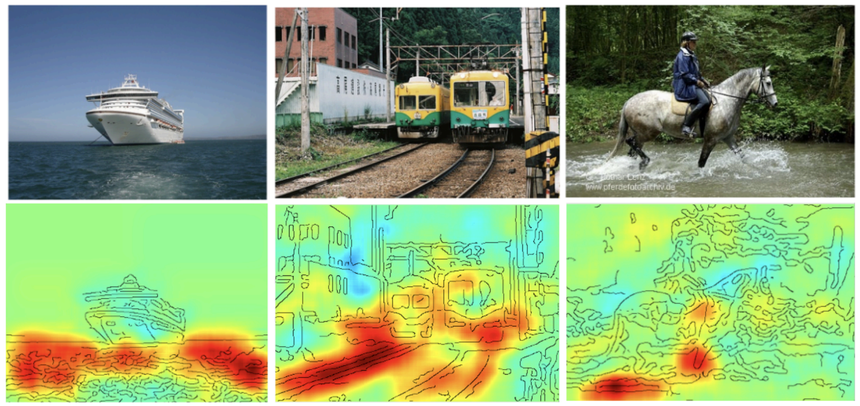
How does the model arrive at this decision?

What can we learn from this?

Explainable AI (XAI) — overview

Local → Global

Heatmaps



Lapuschkin, et al., *Nature communications* 10.1 (2019)
"Unmasking Clever Hans predictors and assessing what machines really learn."

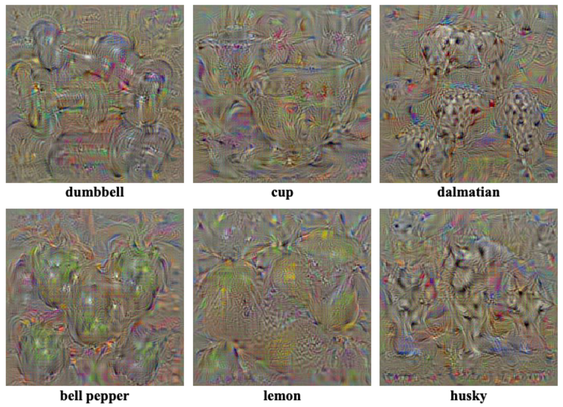
Relevant input regions?

Discover concepts

Vielhaben, et al., *arXiv:2203.06043* (2022).
"Sparse Subspace Clustering for Concept Discovery (SSCCD)."

Investigate hidden model representations

Activation maximization



Simonyan, et al., *arXiv:1312.6034* (2013).
"Deep inside convolutional networks: Visualising image classification models and saliency maps."

Visualize class prototypes

Counterfactuals

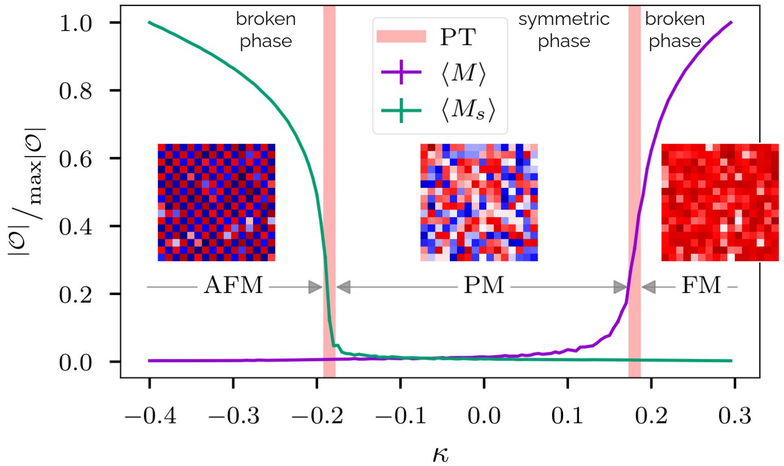
original not blond | adversarial example blond ($p \approx 0.99$) | counterfactual blond ($p \approx 0.99$)

Dombrowski, et al., *arXiv:2206.05075* (2022).
"Diffeomorphic Counterfactuals with Generative Models."

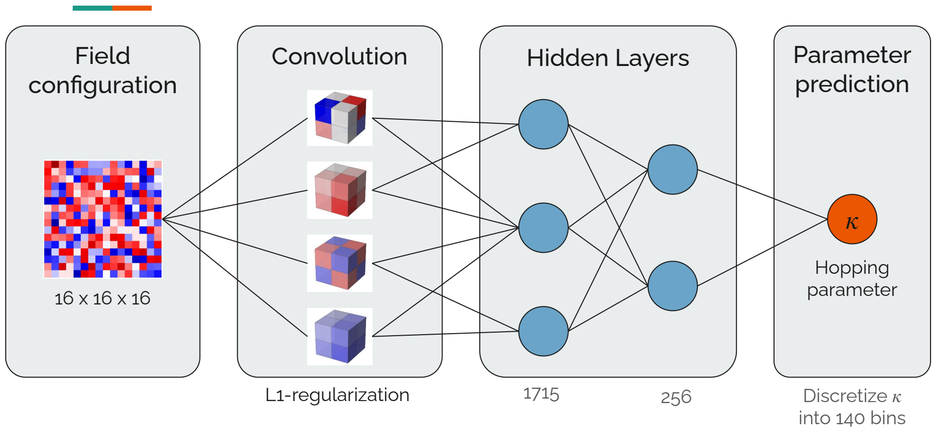
Which input changes prediction?

Use-case: scientific insights with XAI

Yukawa phase diagram - fixed fermion coupling

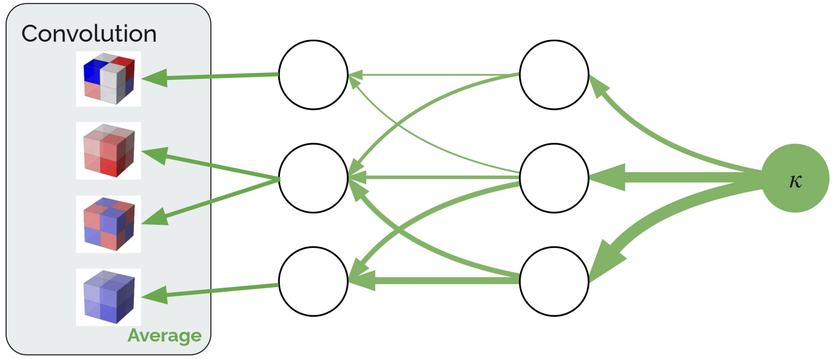


Network architecture - shallow convolution



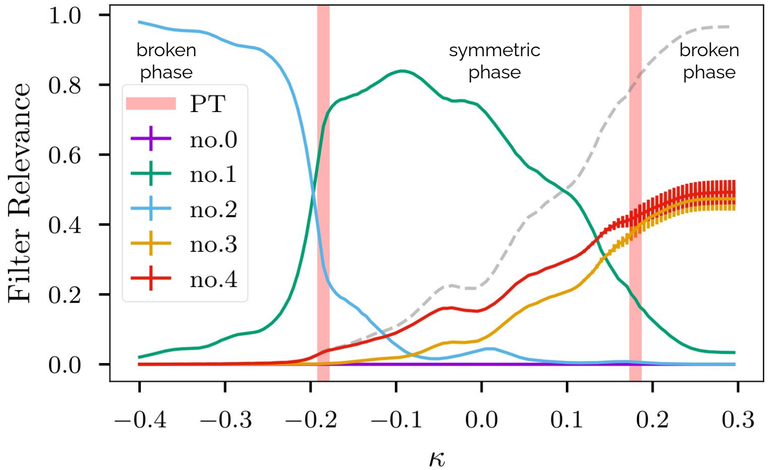
Analysis using XAI

Filter relevance - averaging conv layer



Graphic adapted from *heatmapping.org*
G. Montavano, W. Samek, K.-R. Müller Digital Signal Processing 73 (2018) 1-15

Filter relevances - discovering structures



Blücher, et al., *Physical Review D* 101.9 (2020): 094507.
"Towards novel insights in lattice field theory with explainable machine learning."

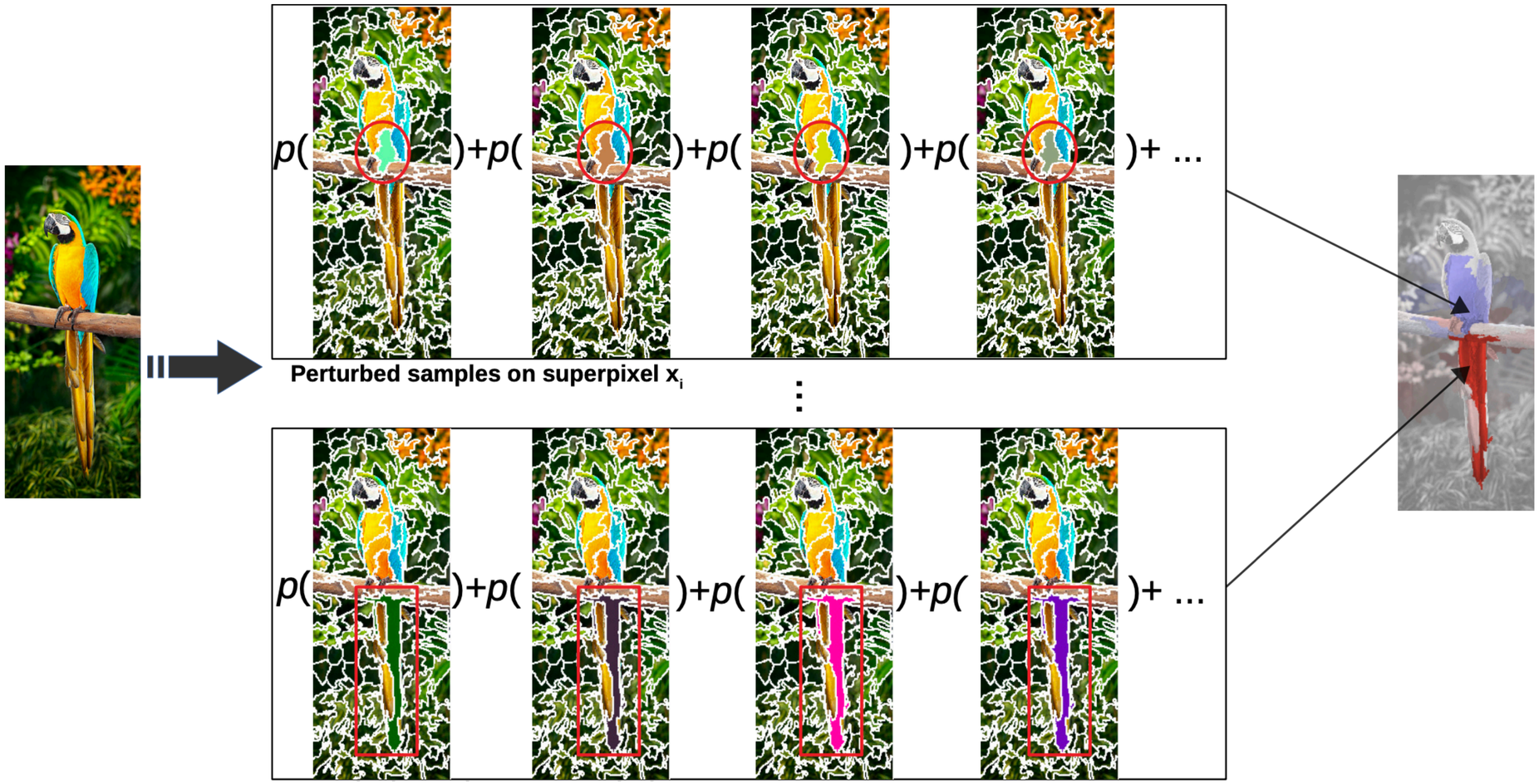
Local model-agnostic attributions

Explaining by removing

Covert, Ian, Scott M. Lundberg, and Su-In Lee.
"Explaining by Removing: A Unified Framework for Model Explanation."
J. Mach. Learn. Res. 22 (2021): 209-1.

Prediction Difference Analysis (*PredDiff*)

Original — Perturbations = Attribution



Wei, Yi, et al., „Explain black-box image classifications using superpixel-based interpretation“. *24th International Conference on Pattern Recognition (ICPR), 2018.*

Prediction Difference Analysis (*PredDiff*)

- Simple attribution framework
- Firmly rooted in probability theory
- Only occlude target feature Y

  —> numerically cheap

  —> on-manifold

Occluded prediction

Model

Imputer

$$m^f_{Y|x_i} = \int \mathrm{d}Y f(x, Y) p(Y \mid x)$$

Regression

$$\bar{m}^f_{Y|x_i} := f(x_i, y_i) - m^f_{Y|x_i}$$

Classification

$$\bar{m}^{f_c}_{Y|x_i} := \log_2 f_c(x_i, y_i) - \log_2 m^{f_c}_{Y|x_i}$$

Robnik-Šikonja, et al.
"Explaining classifications for individual instances."
IEEE Transactions on Knowledge and Data Engineering 20.5 (2008): 589-600.

Blücher, et al. *Artificial Intelligence* 312 (2022): 103774.
"PredDiff: Explanations and interactions from conditional expectations."

Shapley values — a „n-player game“

Marginalize spectator features

Aggregate coalitions S

Normalization ($|\mathcal{X}| = n$)

e.g. $v_x(S) = m_{X_{\bar{S}}|x_S}$

Value function (game)
$v : \{0,1\}^n \mapsto \mathbb{R}$

$$\phi_j(v) = \sum_{S \subseteq \mathcal{X} \setminus \{X^j\}} \frac{|S|!(n - |S| - 1)!}{n!} \left[v(S \cup X^j) - v(S) \right]$$

Marginal contribution

Original | $|S| = 10\,\%$ | $|S| = 25\,\%$ | $|S| = 50\,\%$ | $|S| = 75\,\%$

Cardinality coalitions S

Lundberg, Scott M., and Su-In Lee. *Advances in neural information processing systems* 30 (2017).
"A unified approach to interpreting model predictions."

Shapley values — desirable axioms

Efficiency

$$\sum_i \phi_i = v(\mathcal{X}) - v(\emptyset)$$

Symmetry

$$\forall S \subseteq \mathcal{X} \backslash \{i, j\} : v(S \cup i) = v(S \cup j)$$
$$\Rightarrow \phi_i = \phi_j$$

Dummy player

$$\forall S \subseteq \mathcal{X} \backslash i : v(S \cup i) = v(S)$$
$$\Rightarrow \phi_i = 0$$

Linearity

$$\phi_i(v + \alpha w) = \phi_i(v) + \alpha \phi(w)$$

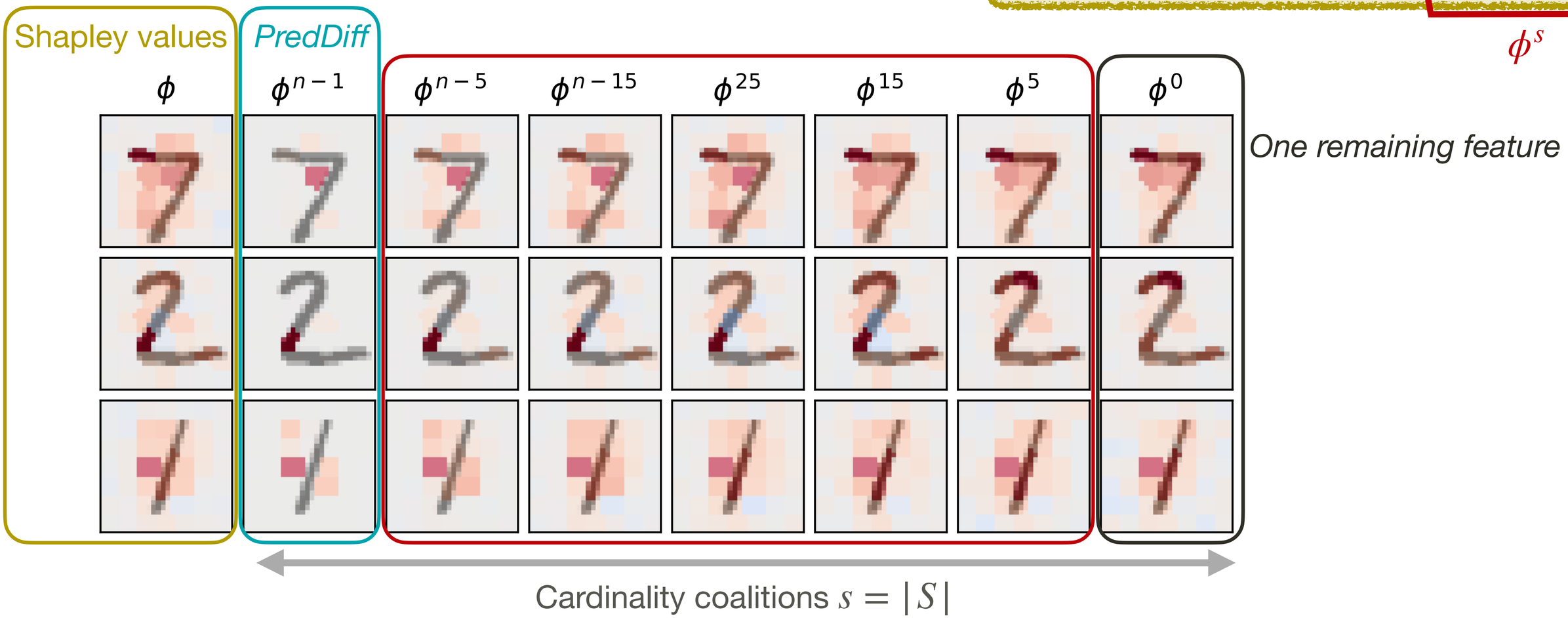
$$\phi_j(v) = \sum_{S \subseteq \mathcal{X} \backslash \{X^j\}} \frac{|S|!(n - |S| - 1)!}{n!} \left[v(S \cup X^j) - v(S) \right]$$

Uniquely distributes
all player's contribution

Lundberg, Scott M., and Su-In Lee. *Advances in neural information processing systems* 30 (2017).
"A unified approach to interpreting model predictions."

Shapley values — marginal contributions

$$\phi_j(v) = \sum_{S \subseteq \mathcal{X} \setminus \{X^j\}} \frac{|S|!(n-|S|-1)!}{n!} \underbrace{\left[v(S \cup X^j) - v(S)\right]}_{\phi^s}$$



PredDiff higher-order attributions

„no-interaction“ property

Regression

$f(X, Y) = Y^2 + 3Z$

Classification

Conditional informative interaction

$p(y, z \mid c, x) = p(y \mid c, x)\ p(z \mid c, x)$

Factorizing imputer distribution

$$\bar{m}^{f^{YZ}}_{YZ|x} = 0$$

Blücher, et al. *Artificial Intelligence* 312 (2022): 103774.
"PredDiff: Explanations and interactions from conditional expectations."

Interaction measure

Decompose model prediction

$$f^V(V) = \sum_{W \subseteq V} (-1)^{v-w} P_{\mathcal{X} \setminus W} f(X, Y, Z)$$

for example $P_Y f(X, Y, Z) = f(X, y, Z)$

Two-point interaction

$$\bar{m}^{f^{YZ}}_{YZ|x} = \bar{m}^{f}_{YZ|x} - \bar{m}^{f^Y}_{Y|x} - \bar{m}^{f^Z}_{Z|x}$$

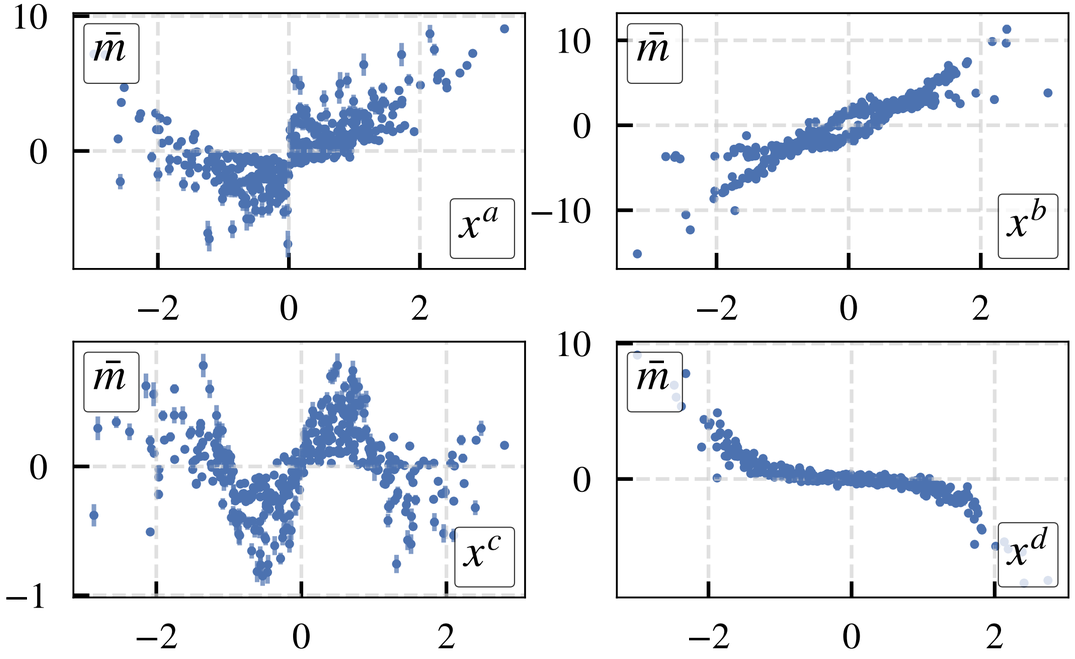
Tsai, et al. *arXiv:2203.00870* (2022).
"Faith-Shap: The Faithful Shapley Shapley Interaction Index."

Experiments

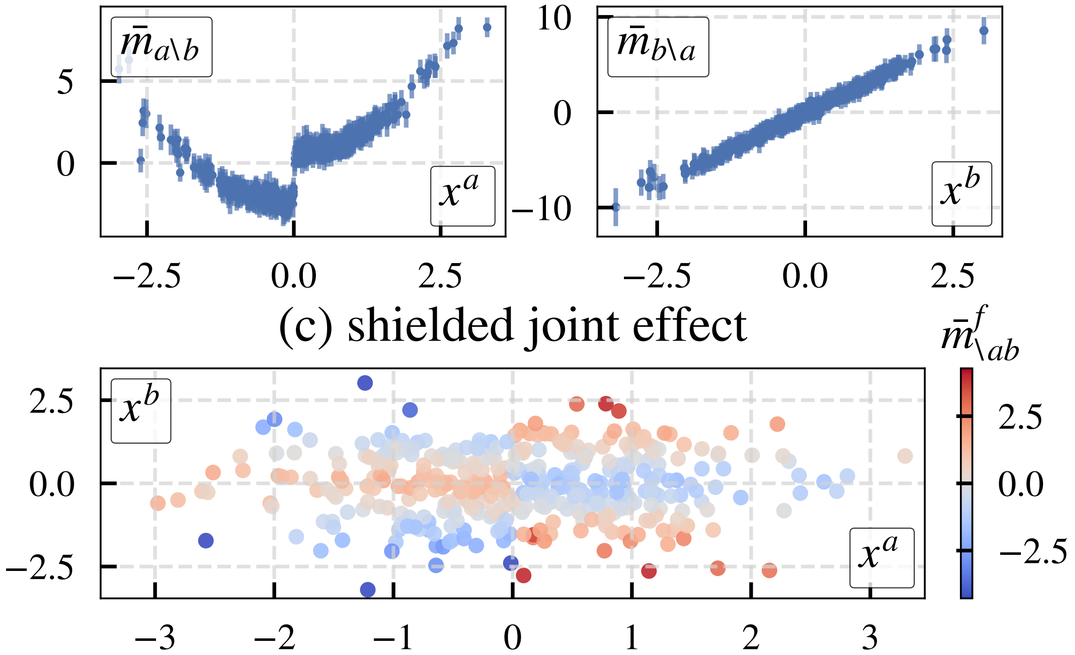
Recover additive regression components

$$f(x) = x_a^2 + 3x_b + \sin(\pi x_c) - \frac{x_d^3}{2} + 2\,\mathrm{sgn}\,(x_a)\,\mathrm{abs}(x_b)$$

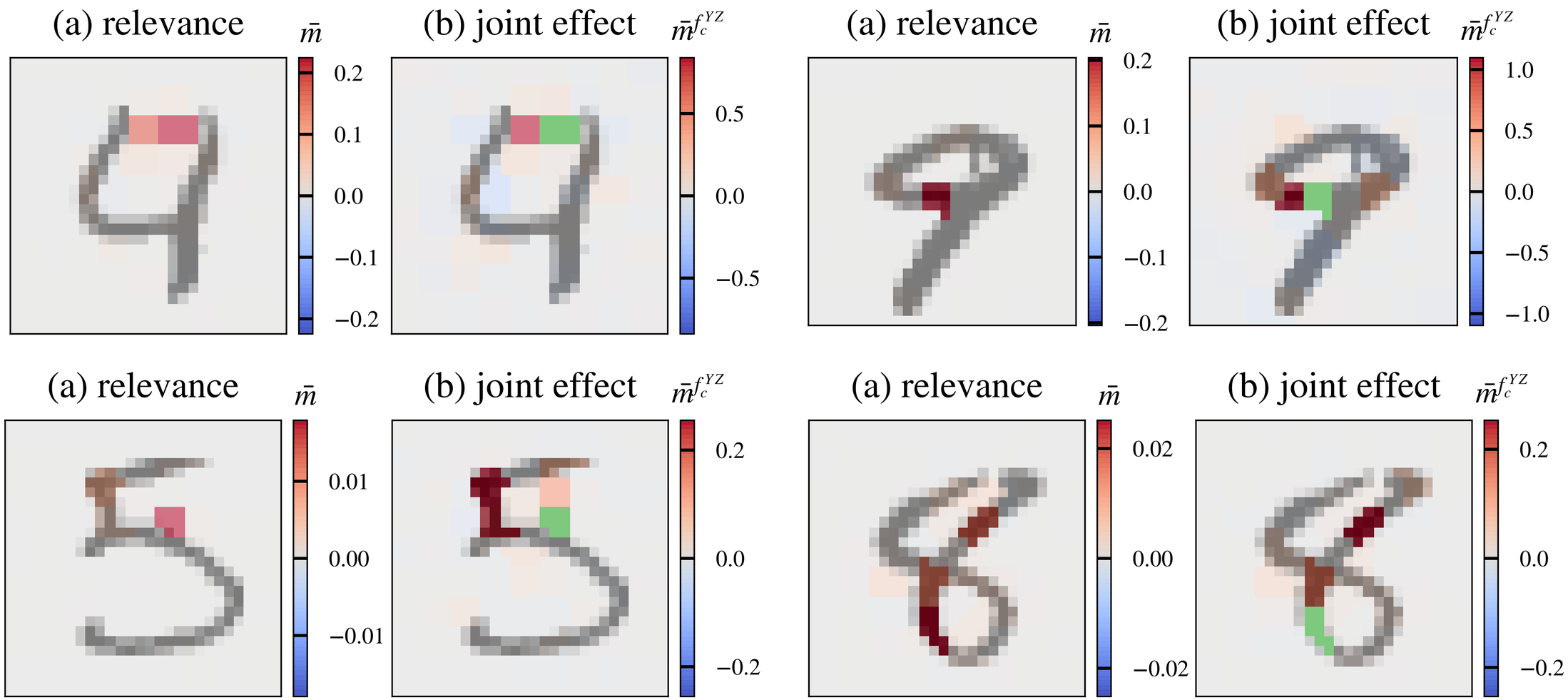
Main effects



Interactions



Meaningful interactions for MNIST digits



reference superpixel

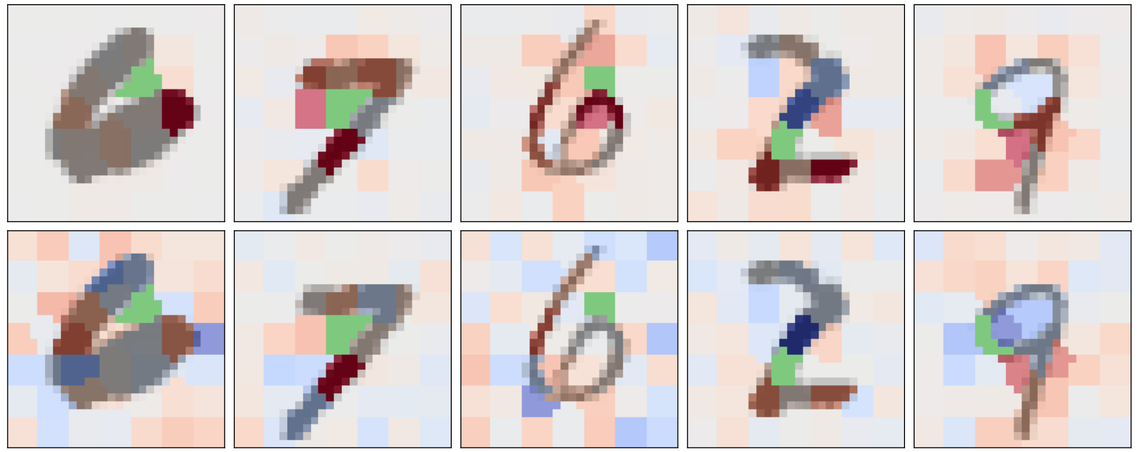
Similar attributions for *PredDiff* & Shapley

Relevances

PredDiff

Shapley values

Interactions

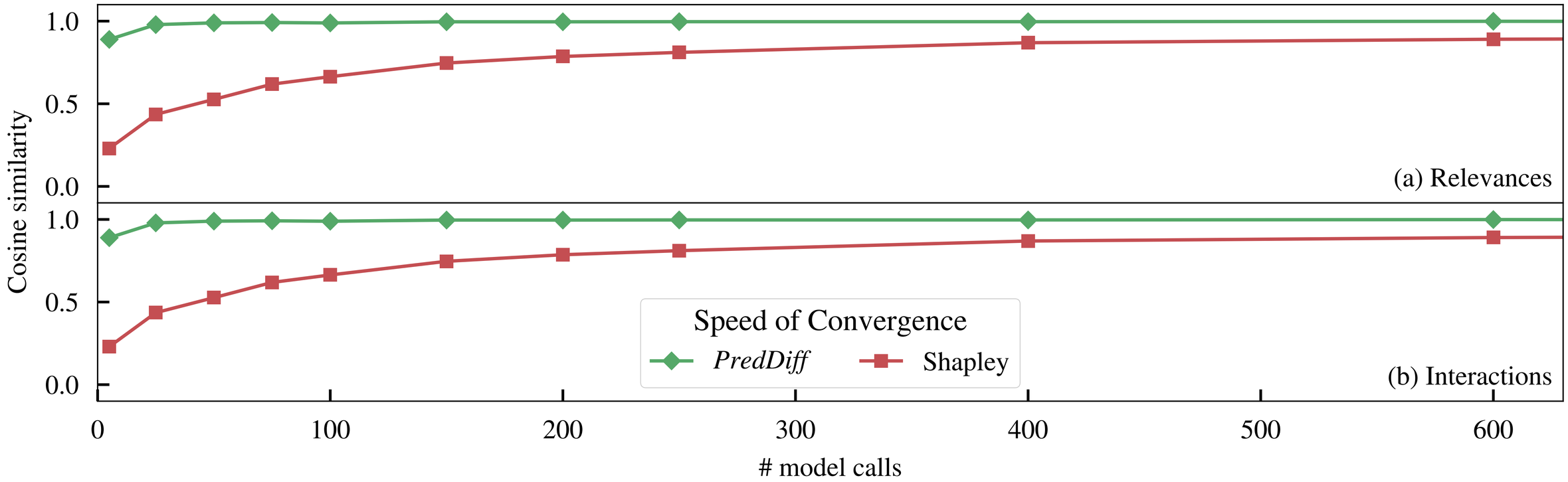


PredDiff

Shapley values

Blücher, et al. *Artificial Intelligence* 312 (2022): 103774.
"PredDiff: Explanations and interactions from conditional expectations."

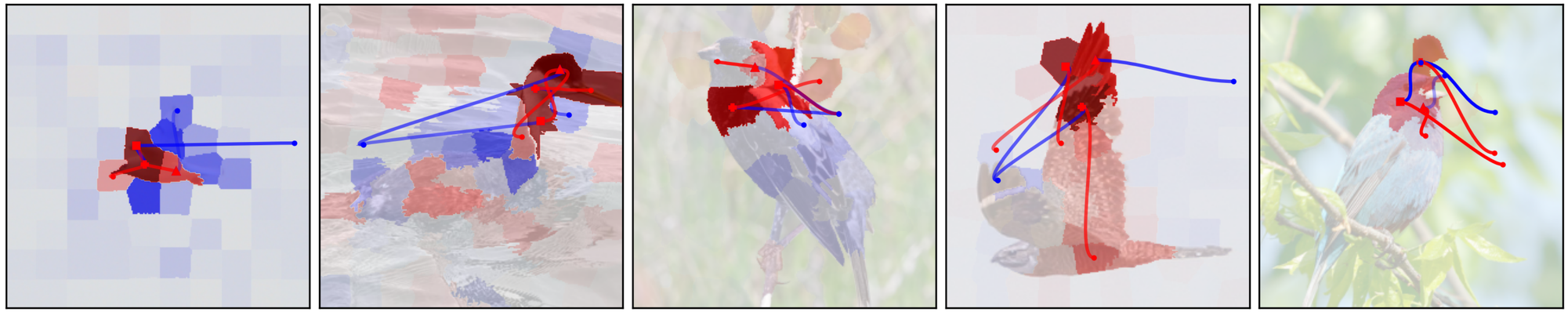
PredDiff: rapid convergence & scalability



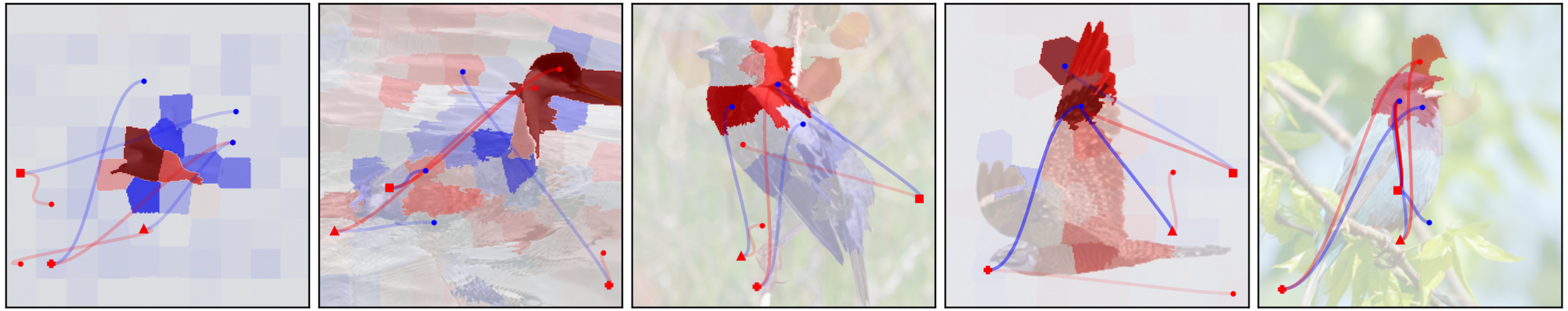
Blücher, et al. *Artificial Intelligence* 312 (2022): 103774.
"PredDiff: Explanations and interactions from conditional expectations."

Insight: relevant features show interactions

Highest/Lowest relevance reference super pixels



Random reference super pixels



Take-home messages

Take-home messages

PredDiff

- Conceptual simplicity
- On-manifold imputations
  only target feature is occluded
- Linear scaling
- Provably *no-interaction* property

Higher-order attributions **NEW**

- More detailed insights through interaction effects
- Allows to understand complex vision and text classifiers
- Software package to appear

Shapley values

- Very popular in the literature
- Axiomatic interpretation
- Many different variants proposed

Useful resources

- heatmapping.org
- https://github.com/chr5tphr/zennit

- https://captum.ai/
- https://github.com/slundberg/shap