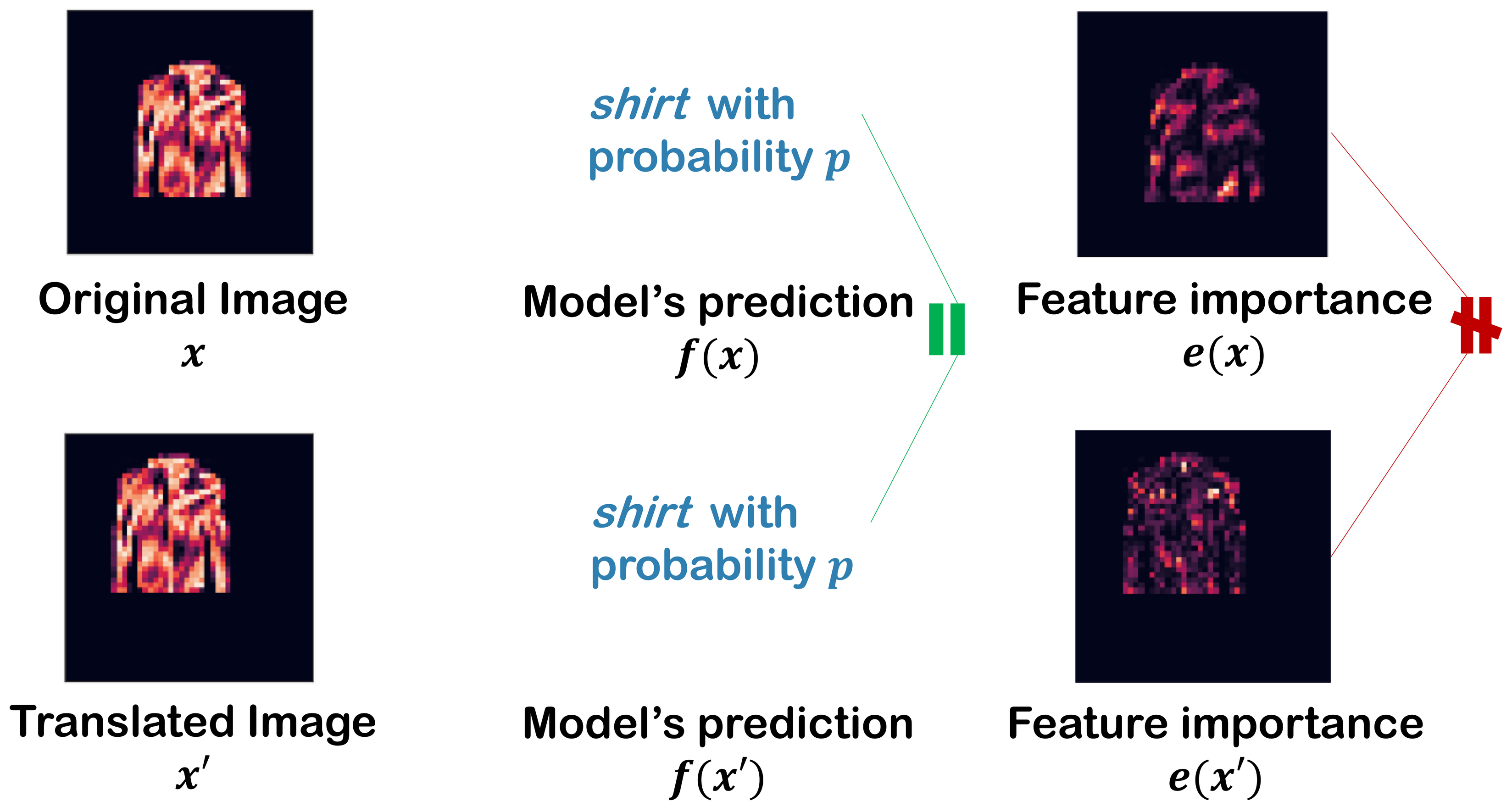
Evaluating the Robustness of Interprerability Methods through Explanation Invariance and Equivariance

van_der_Schaar \ LAB

UNIVERSITY OF CAMBRIDGE

Jonathan Crabbé and Mihaela van der Schaar

1. Misleading Explanations



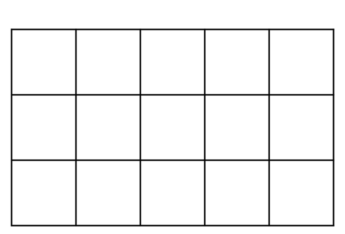
Many ML models are robust to input symmetries (e.g. CNNs with translations, GNNs with node permutation).

If a model's predction does not change by applying a symmetry to its input (invariance), we expect the same for the explanations.

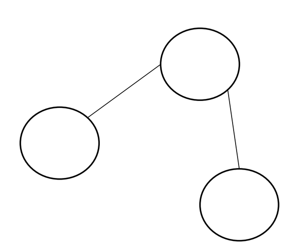
Our first finding is that many popular interpretability methods (e.g. GradSHAP, TCAV) do not always verify this desideratum.

2. Geometric Deep Learning Concepts

Domains. It is the support Ω on which data is defined

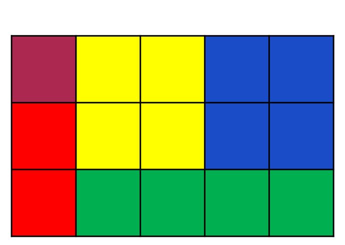


Images: $\Omega = \mathbb{Z}_H \times \mathbb{Z}_W$

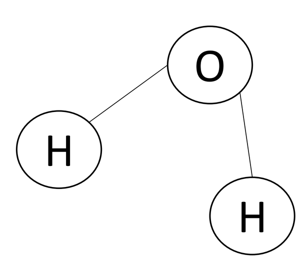


Graph data: $\Omega = (\mathcal{V}, \mathcal{E})$

Signals. A signal is a function $x: \Omega \to \mathcal{C}$ mapping the domain to a vector space $\mathcal{C}$

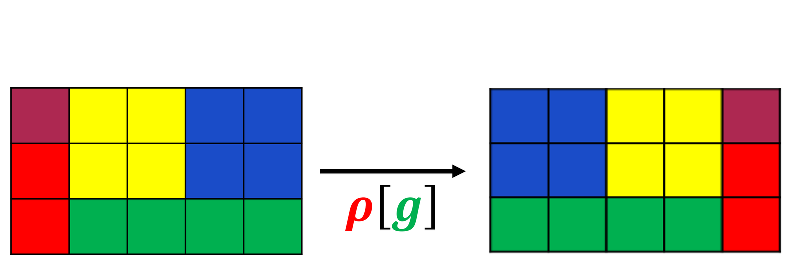


RGB Images: $\mathcal{C} = \mathbb{R}^3$

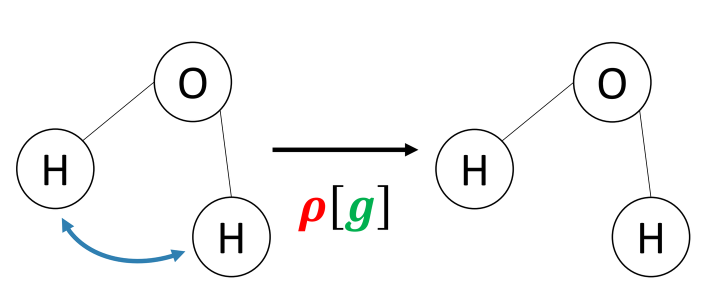


Molecules: $\mathcal{C} = \mathbb{R}^{N_{atoms}} \oplus \mathbb{R}^{N_{valence}}$

Symmetry group. It is a set $\mathcal{G}$ of transformations preserving a signal information. Each symmetry $g \in \mathcal{G}$ acts on x via a representation $\rho[g] \in \mathbb{R}^{d_X \times d_X} : x' = \rho[g]x$



Mirror symmetries



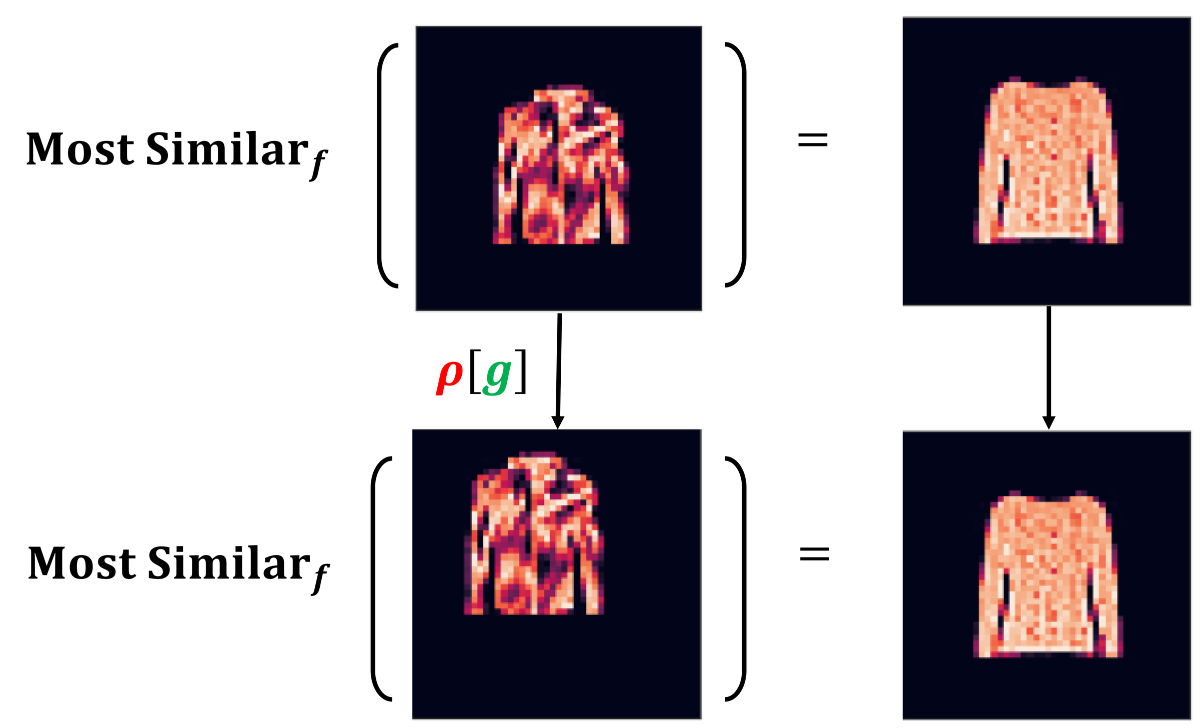
Node permutation

3. Explanation Invariance & Equivariance

Consider an explanation $e: \mathbb{R}^{d_X} \to \mathbb{R}^{d_E}$ for a $\mathcal{G}$-invariant model $f: \mathbb{R}^{d_X} \to \mathbb{R}^{d_Y}$. We distinguish 2 different behaviours for explanations under the symmetry group.

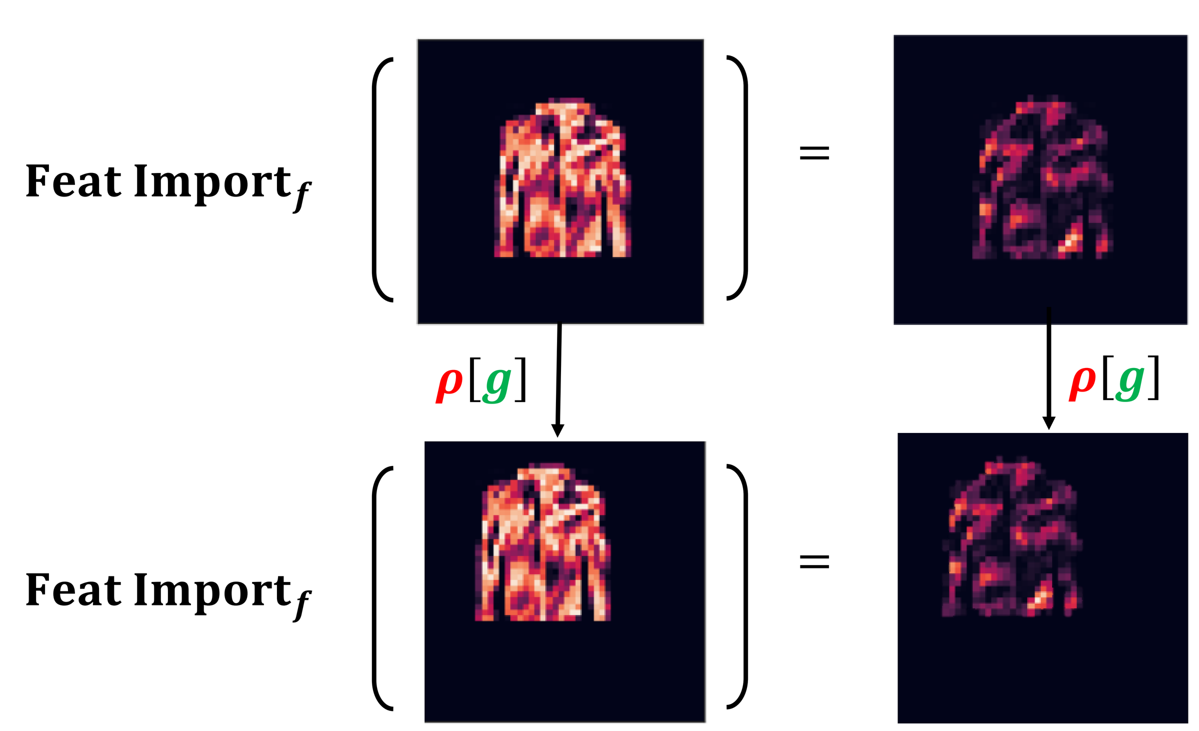
Invariant explanations are unaffected by group symmetries

$$e(\rho[g]x) = e(x)$$



Equivariant explanations transform as the input

$$e(\rho[g]x) = \rho[g]e(x)$$



Note that these prescription apply to other interpretability methods (e.g. it makes sense for concept-based explanation to be invariant and for counterfactual explanations to be equivariant).

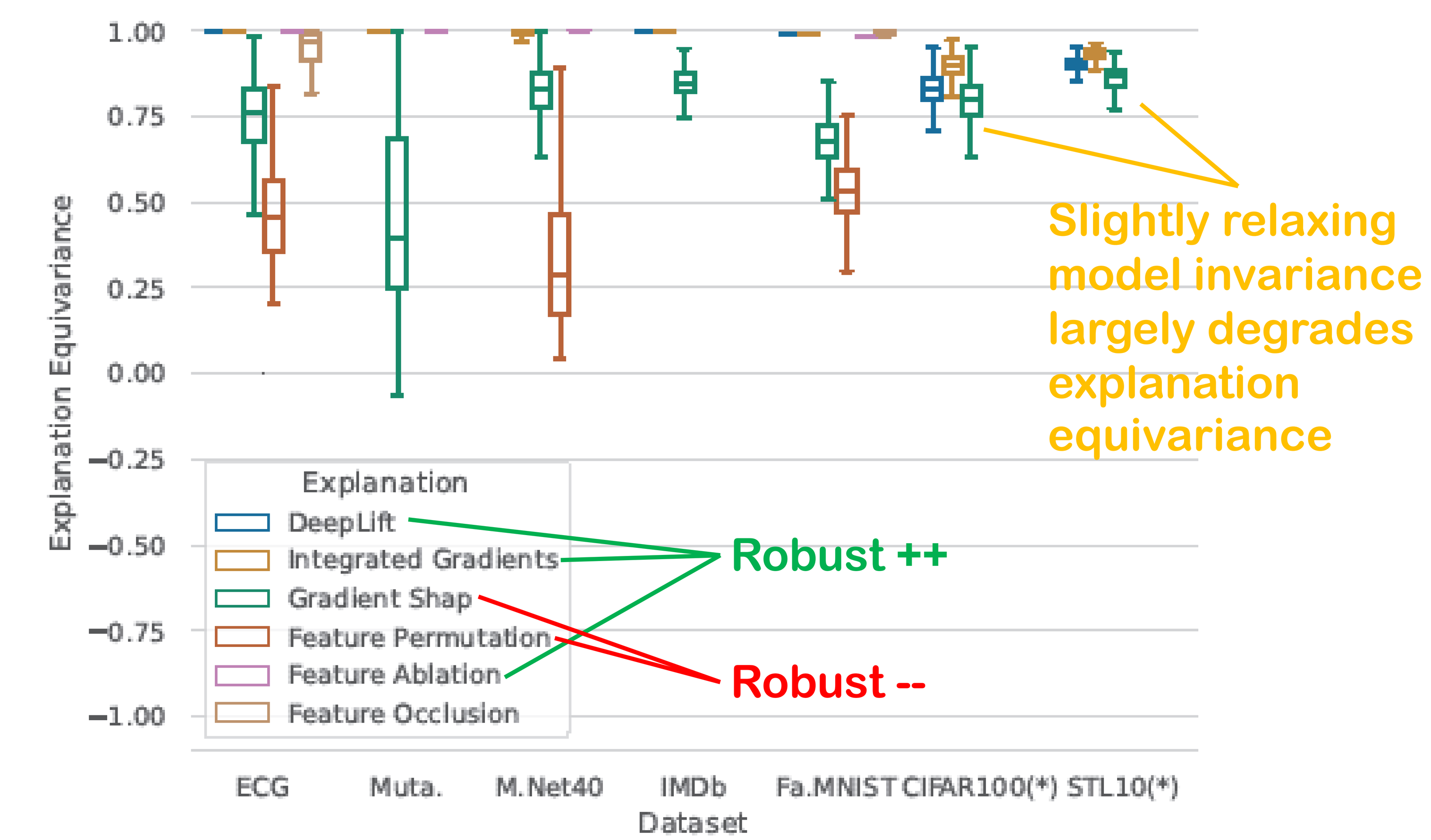
We introduce two metrics to measure to what extent these properties are verified

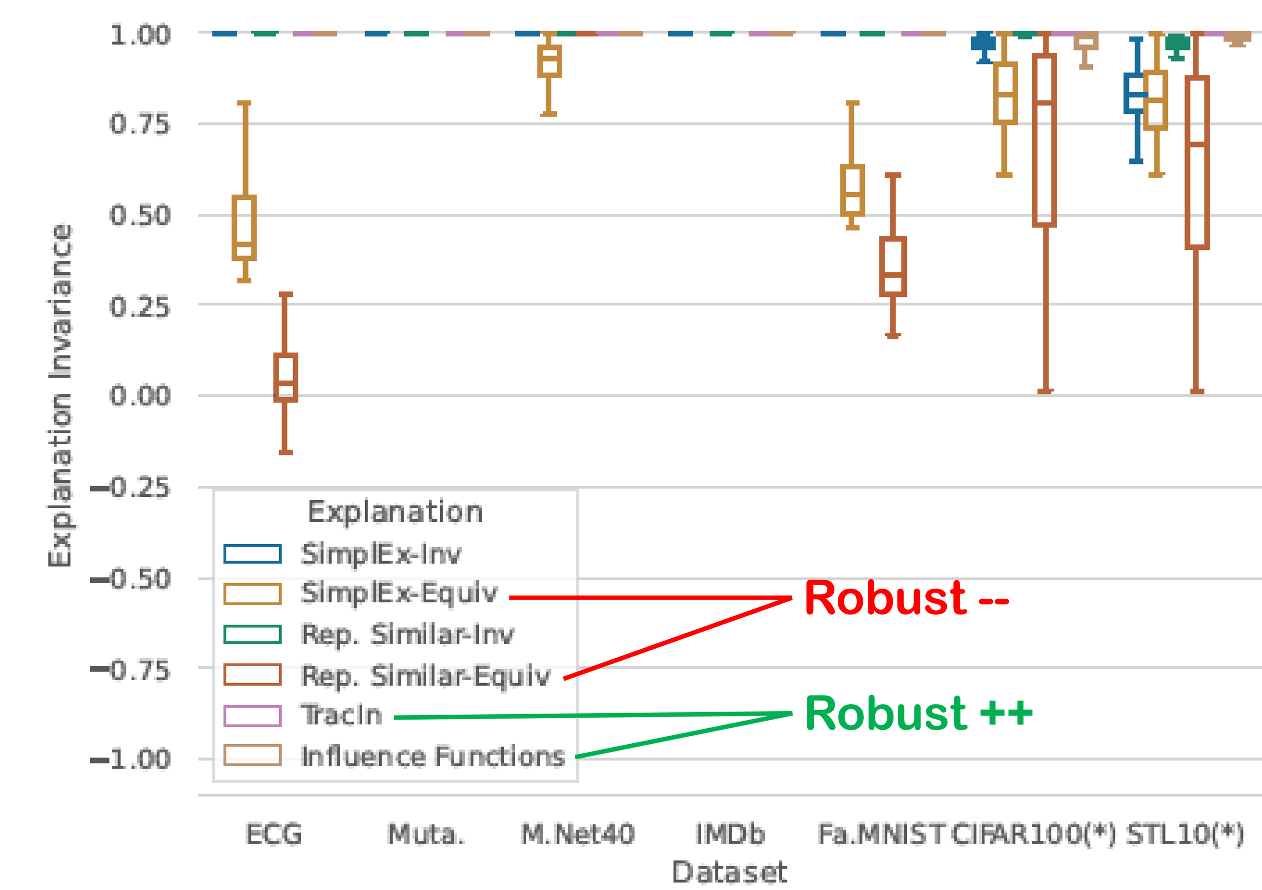
$$\mathrm{Inv}_{\mathcal{G}}[e,x] = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \cos[e(\rho[g]x), e(x)] \quad \in [-1,1]$$

$$\mathrm{Equiv}_{\mathcal{G}}[e,x] = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} \cos[e(\rho[g]x), \rho[g]e(x)] \quad \in [-1,1]$$

NB. These metrics are typically aggregated over several x.

4. Empirical Results





With an empirical analysis on various datasets/modalities/symmetry groups, we observe that some methods are consistently better.

A theoretical analysis explains these differences (e.g. gradient-based methods require invariant baselines).

We provide a flowchart to guarantee explanations that are robust to symmetries.

5. More Information

The paper

https://arxiv.org/abs/2304.06715

My website

https:// jonathancrabbe.github.io