Explaining What Matters: Faithfulness in Molecular Deep Learning

Marcel Hiltscher, Marc Bianciotto, Francesca Grisoni

AICHEMIST SCHOOL, COPENHAGEN 2026

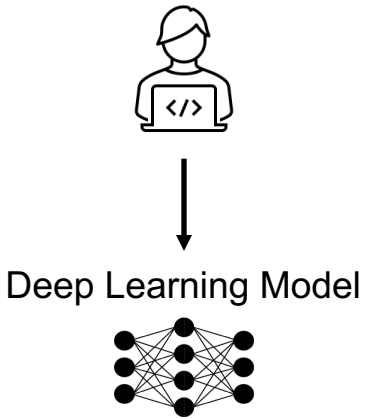
Marcel Hiltscher, 16.06.2026

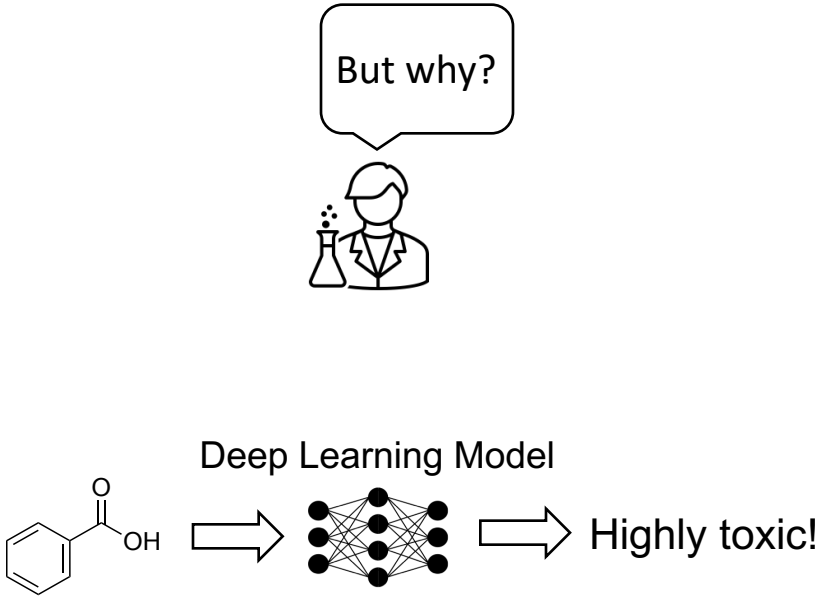
Why explainable AI?

Develop model for toxicity prediction

Use model to evaluate molecules

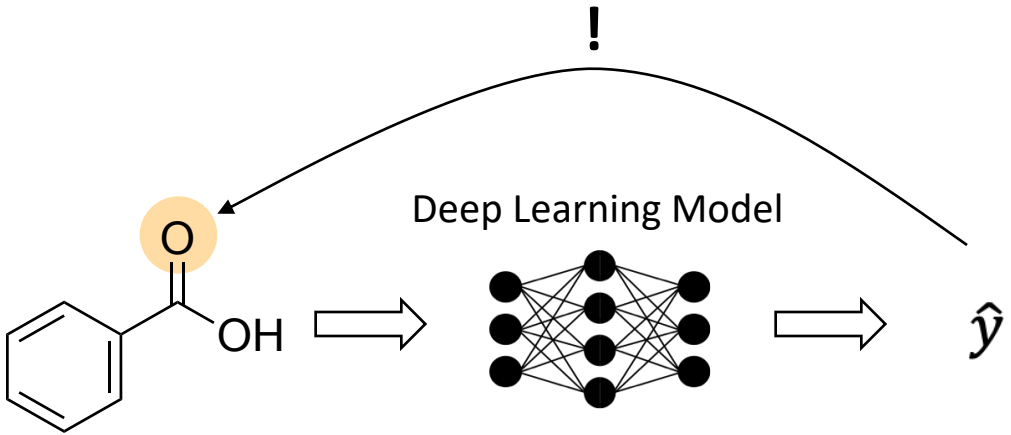
Why explainable AI?





What is explainable AI?

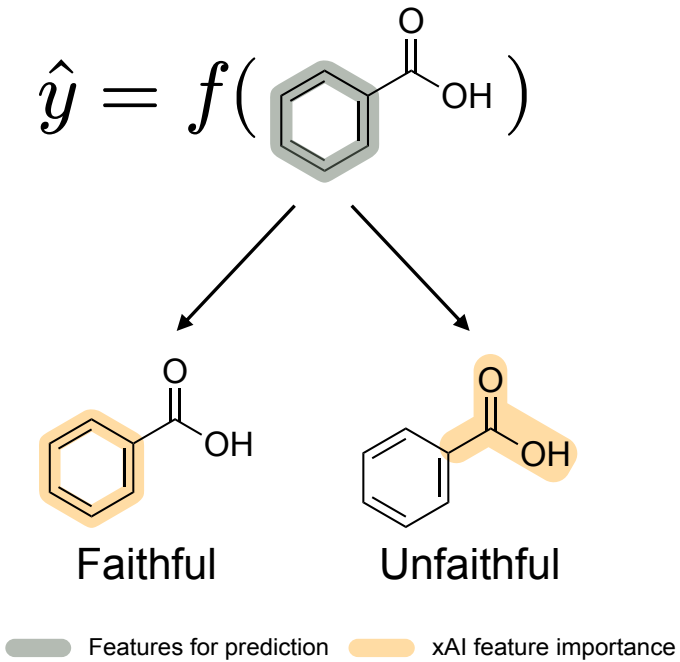
Explainable AI (xAI) aims to make the decisions process of a "black box" model more transparent and understandable



Feature attribution methods

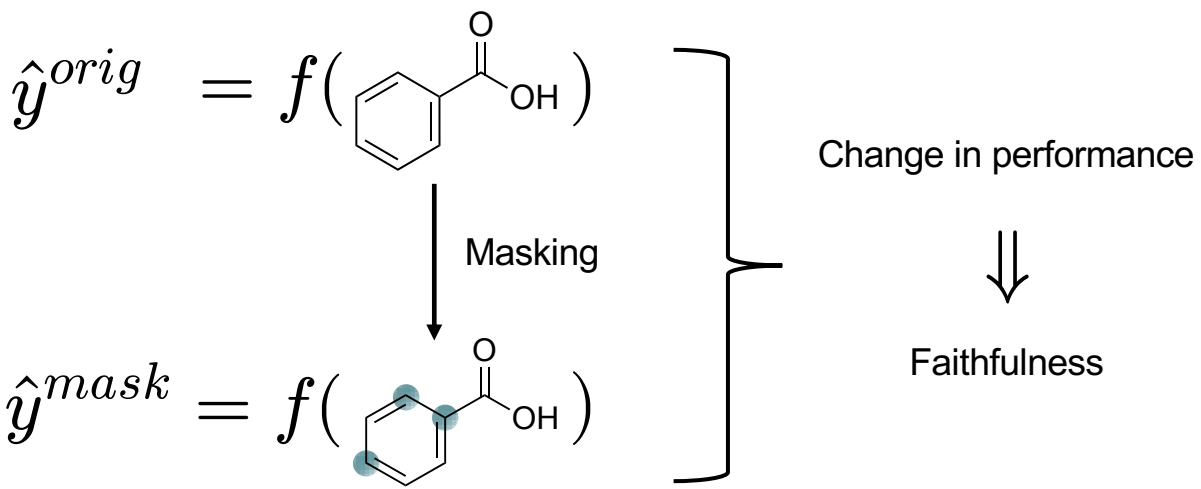
Faithfulness of xAI methods

Are the features that the xAI method highlights indeed important to the model?



Faithfulness via fidelity metric

How do important features identified by the xAI method affect the model prediction?



Zheng, X., Shirani, F., Chen, Z., Lin, C., Cheng, W., Guo, W., & Luo, D. (2025). F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI. In Y. Yue, A. Garg, N. Peng, F. Sha, & R. Yu (Eds), International Conference on Representation Learning (Vol. 2025)

How do we calculate fidelity?

1) Finetune model on masked inputs

→ Make evaluation step robust against OOD

2) Evaluate xAI method for each input via:

$$F\text{-}Fid_i^{+} = |y_i - \hat{y}_i^{\text{mask}}| - |y_i - \hat{y}_i^{\text{orig}}|$$

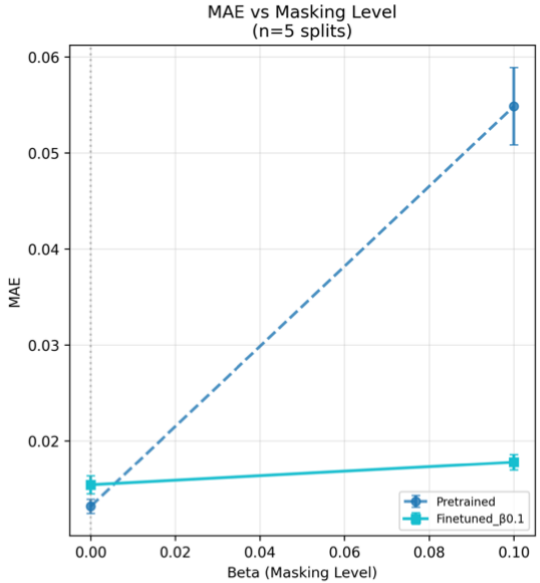
Take an average of 50 different masks

Zheng, X., Shirani, F., Chen, Z., Lin, C., Cheng, W., Guo, W., & Luo, D. (2025). F-Fidelity: A Robust Framework for Faithfulness Evaluation of Explainable AI. In Y. Yue, A. Garg, N. Peng, F. Sha, & R. Yu (Eds), International Conference on Representation Learning (Vol. 2025)

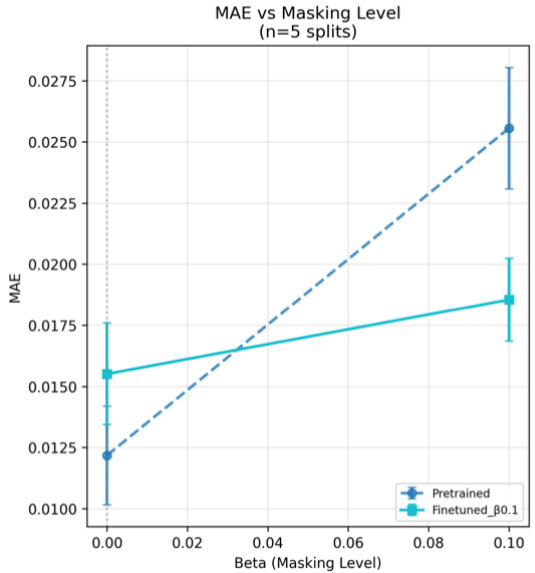
Problem: Different Error Scales

$$F\text{-}Fid_i^+ = |y_i - \hat{y}_i^{\text{mask}}| - |y_i - \hat{y}_i^{\text{orig}}|$$

CNN



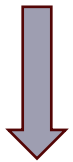
Transformer



F-Fidelity Normalization

Use of random baseline: $F\text{-}Fid^{+}_{i,rand}$

$$Fid^{+}_{i} = \frac{F\text{-}Fid^{+}_{i} - F\text{-}Fid^{+}_{i,rand}}{|F\text{-}Fid^{+}_{i}| + |F\text{-}Fid^{+}_{i,rand}| + \varepsilon}$$

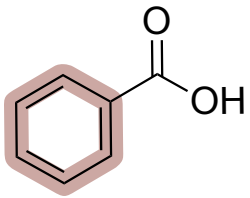


Is the fidelity of the xAI method similar/better/worse than a random baseline?

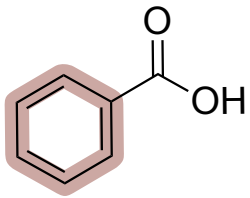
Study Setup: 3 Synthetic Datasets

14k molecules each:

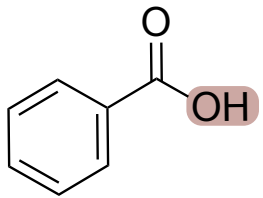
Number of benzene rings



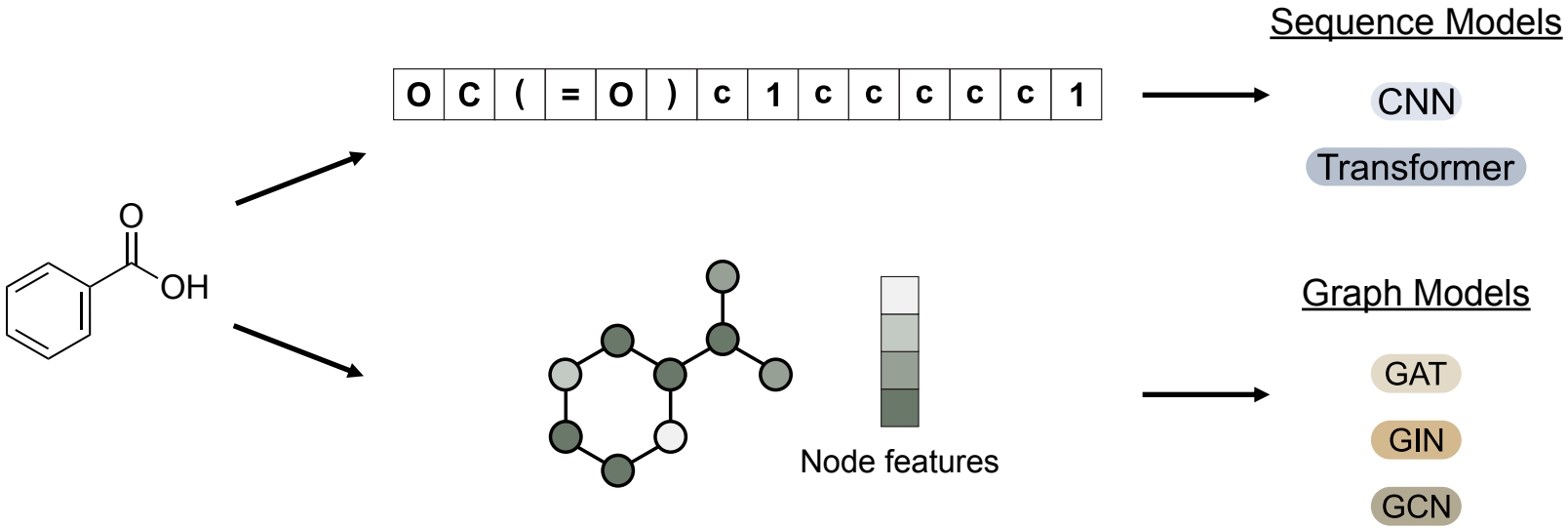
Size of largest conjugated system



Number of H-Bond donors



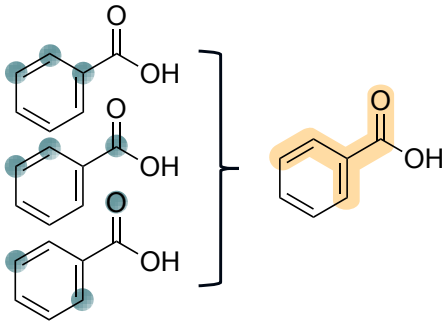
Study Setup: 5 Model Architectures



CNN = Convultion Neural Network GAT = Graph Attention Network GIN = Graph Isomorph Network GCN = Graph Convolution Network

Study Setup: 8 xAI methods

Perturbation Methods



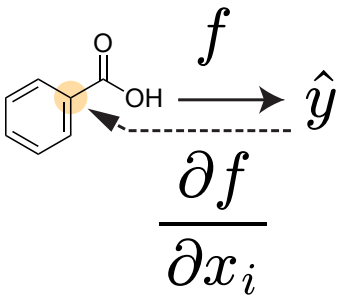
SHAP

KernelSHAP

GraphMASK

GNNExplainer

Gradient Methods

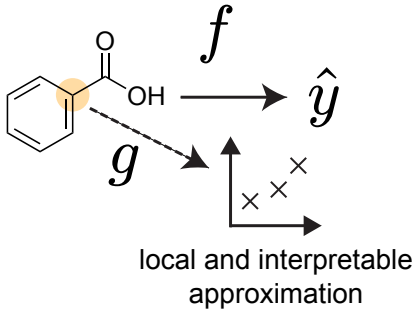


Integrated Gradients (IG)

GradCAM

Guided GradCAM

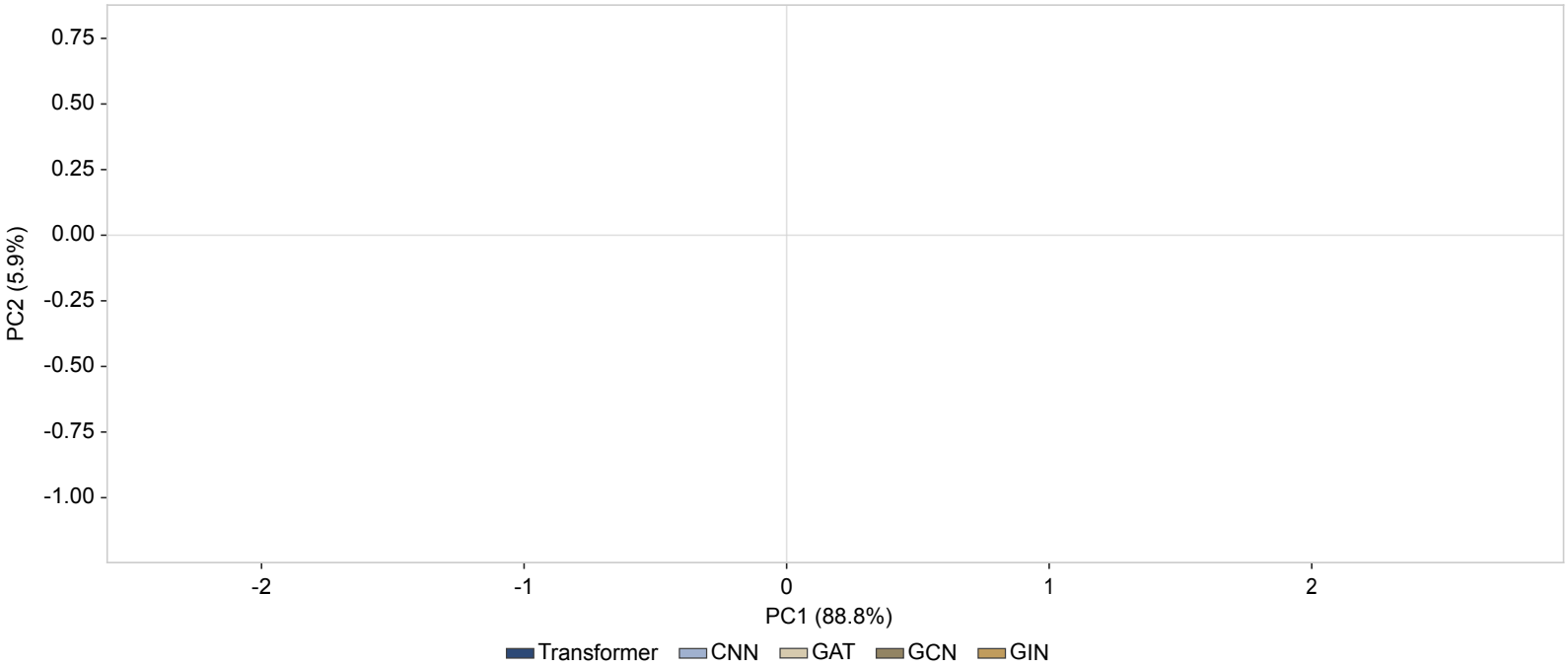
Surrogate Methods



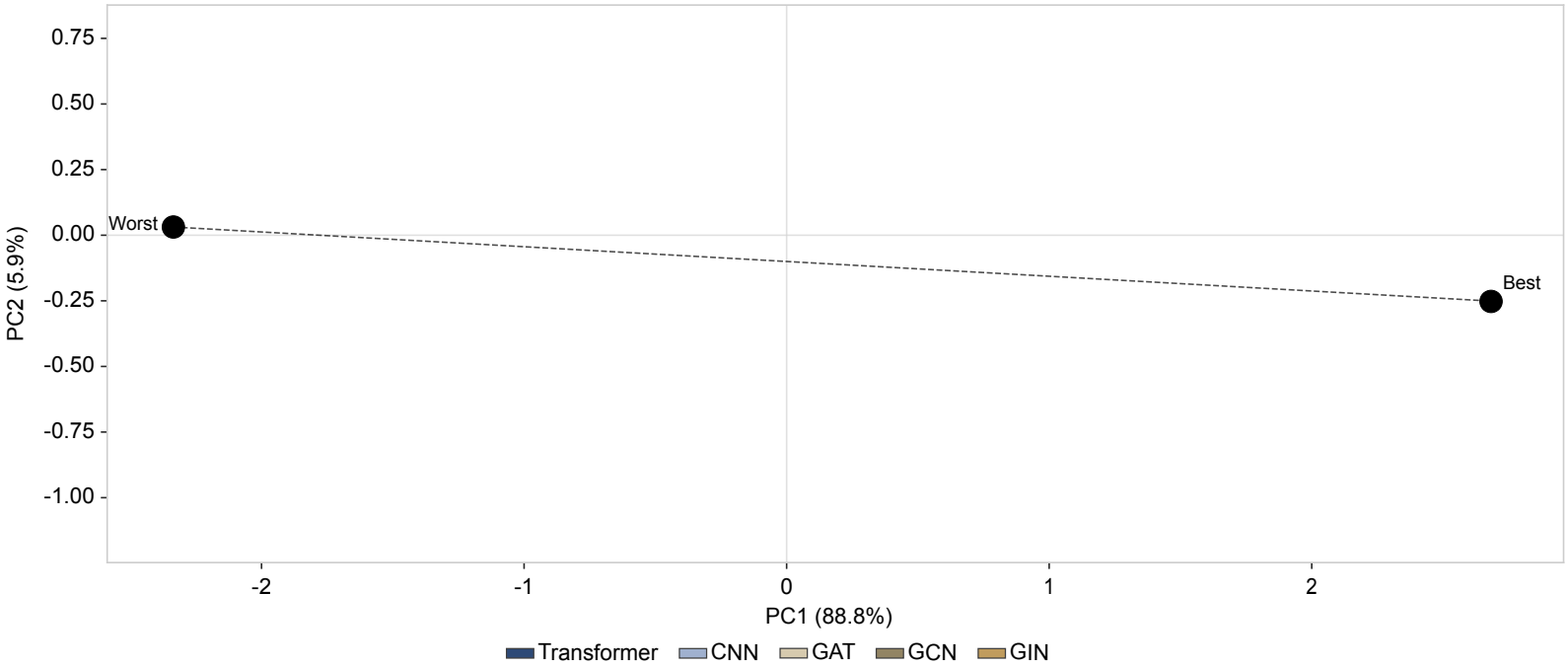
LIME

Masking/Perturbation xAI feature importance

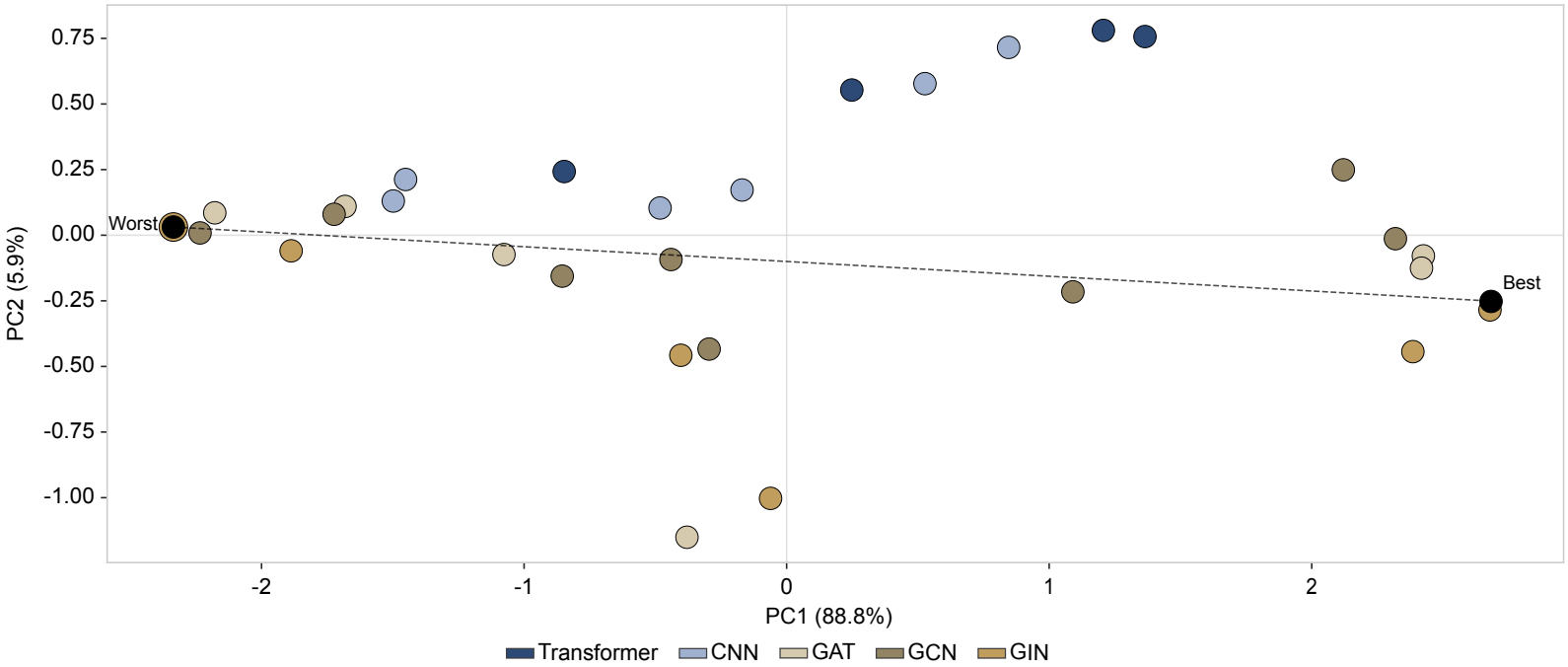
Certain xAI methods offer higher explanation fidelity



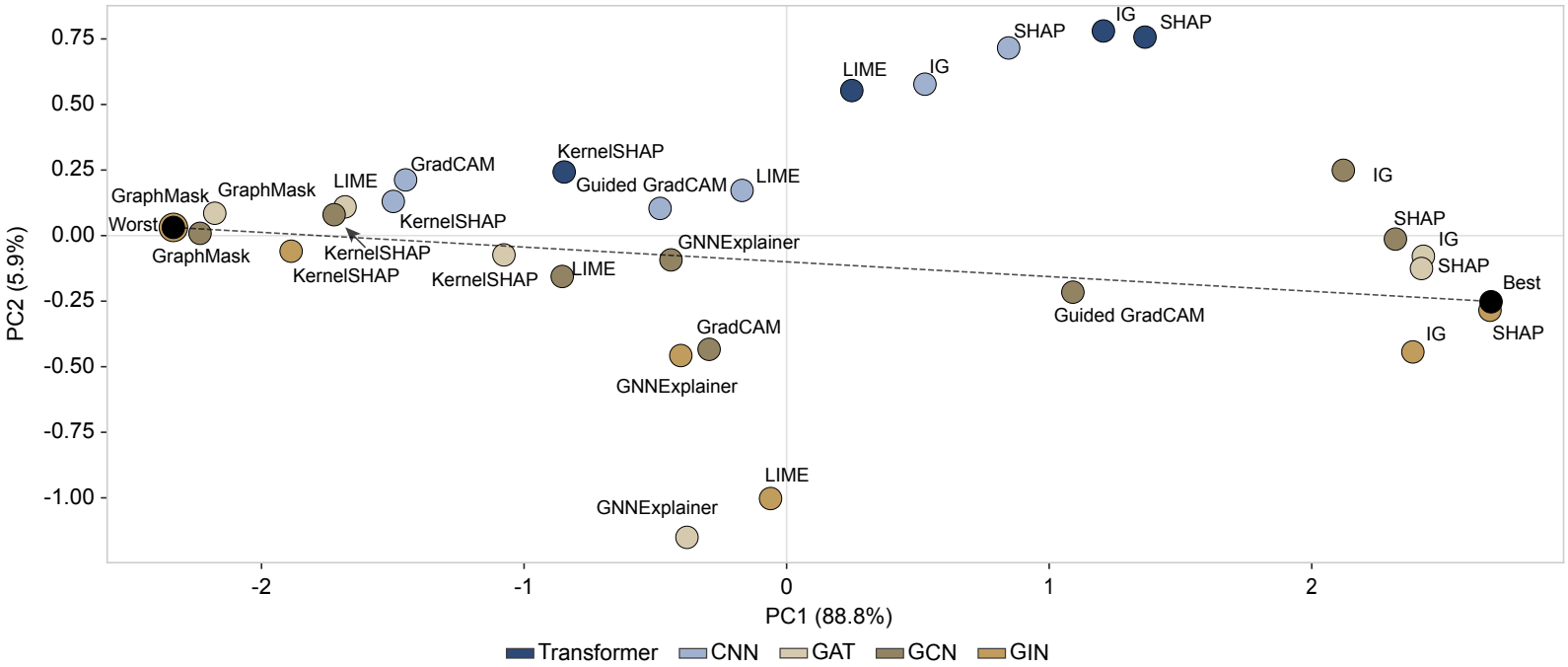
Certain xAI methods offer higher explanation fidelity



Certain xAI methods offer higher explanation fidelity

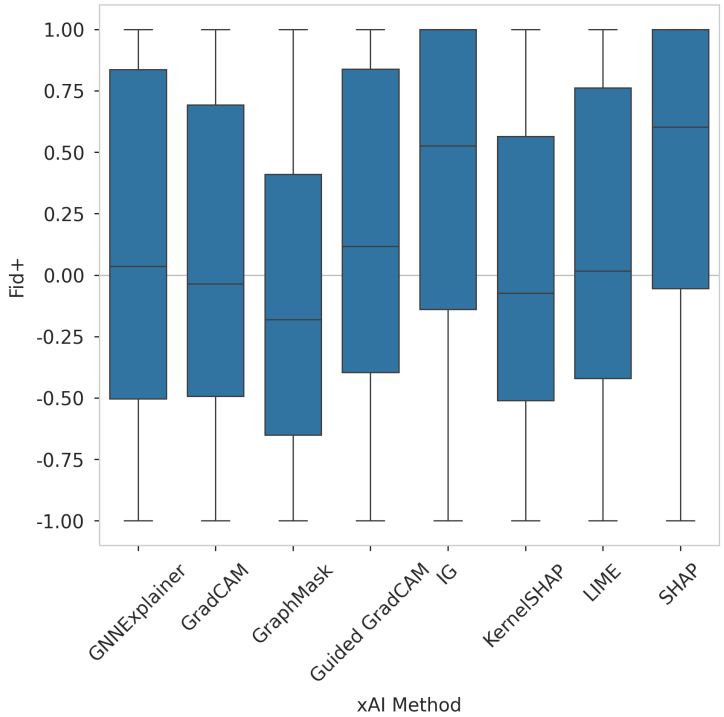


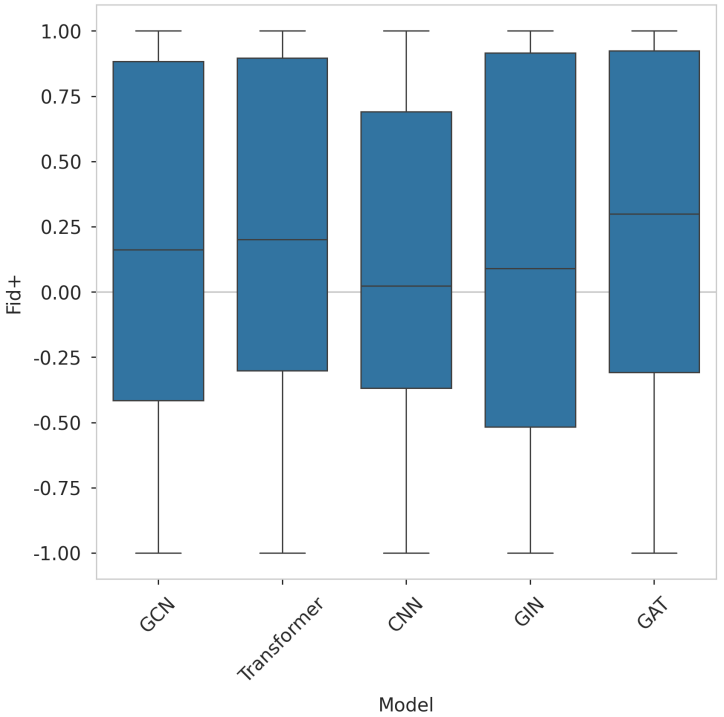
Certain xAI methods offer higher explanation fidelity



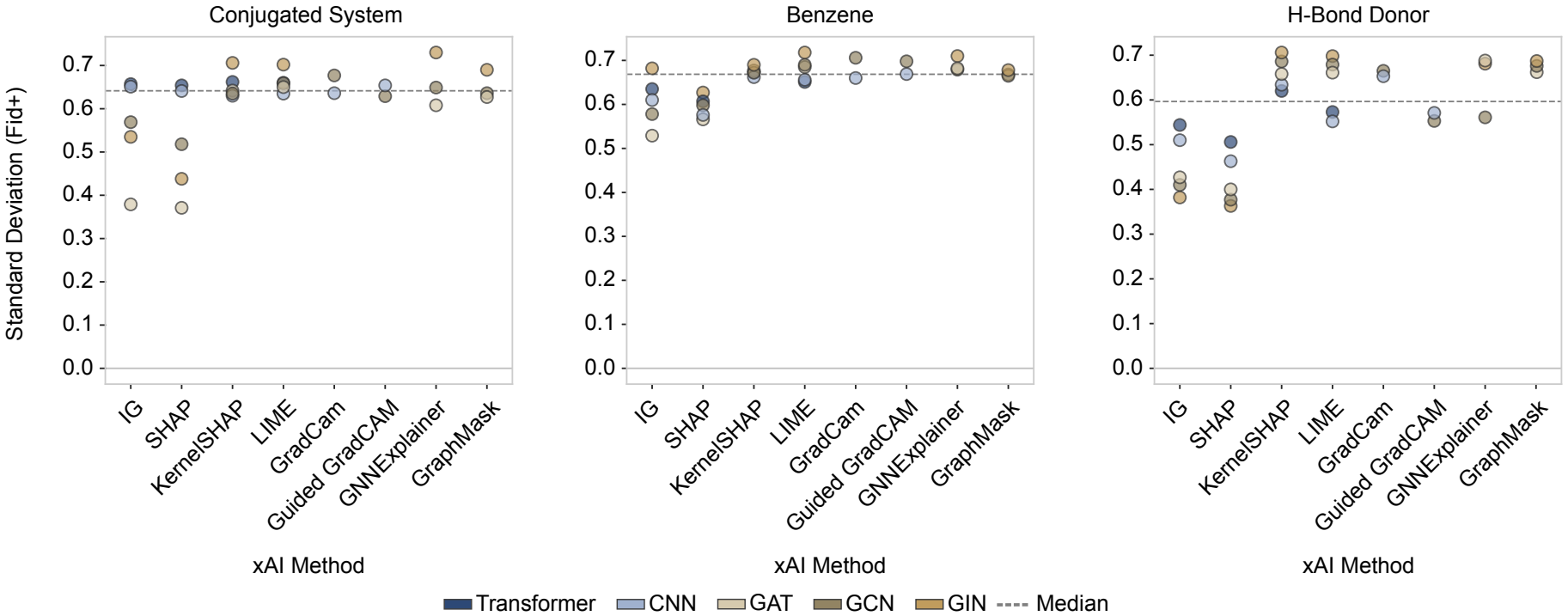
Similar trends in non-synthetic task

Based on the original LogP task



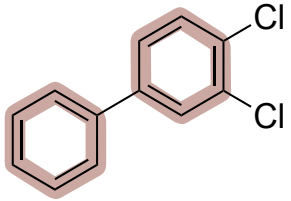


Faithfulness varies a lot per molecule

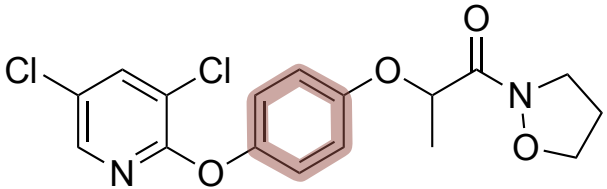


Using the synthetic dataset as evaluation

Synthetic ground-truth (GT) alignment = How well does the xAI method highlight the GT atoms?



GT alignment = 0.90 Fid+ = -1



GT alignment = 0.06 Fid+ = 1

GT Alignment $\in [0, 1]$ Fid+ $\in [-1, 1]$

Synthetic GT Alignment != Faithful

	Conjugated System	Benzene	H-Bond Donor
GNNExplainer	0.01	-0.0	-0.02
GradCAM	-0.02	-0.02	0.15*
GraphMASK	0.0	0.0	0.01
Guided GradCAM	0.03*	-0.04*	0.19*
IG	-0.06*	-0.01	0.05*
KernelSHAP	0.01	0.04*	0.15*
LIME	-0.05*	0.04*	0.09*
SHAP	-0.12*	0.02*	0.04*

Kendall Tau corrleation GT alignment and Fid+

No correlation across all xAI methods and targets!

Summary

1) Some xAI methods provide more faithful model explanations

2) More faithful methods are more robust on graph models

3) Faithfulness varies strongly across inputs
   - Even for more faithful xAI methods

4) Fidelity offers a direct measure of faithfulness
   - GT alignment is not a reliable proxy for faithfulness

Thanks for listening!

Preprint

TU/e
- Francesca Grisoni
- Whole MolML team <3

Sanofi
- Marc Bianciotto

Bayer
- Andi Hunklinger

This study was partially funded by the Horizon Europe funding programme, under the Marie Skłodowska-Curie ActionsDoctoral Networks grant agreement “Explainable AI for Molecules - AiChemist” no. 101120466. It was co-funded by the the European Union (ERC, ReMINDER, 101077879, F.G) This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-13237. M.H. and M.B. are Sanofi employees and may hold shares and/or stocks options in the company