Evaluating the Faithfulness of Fingerprint-based Explanations

Marcel Hiltscher

Table of Contents

What is Faithfulness?

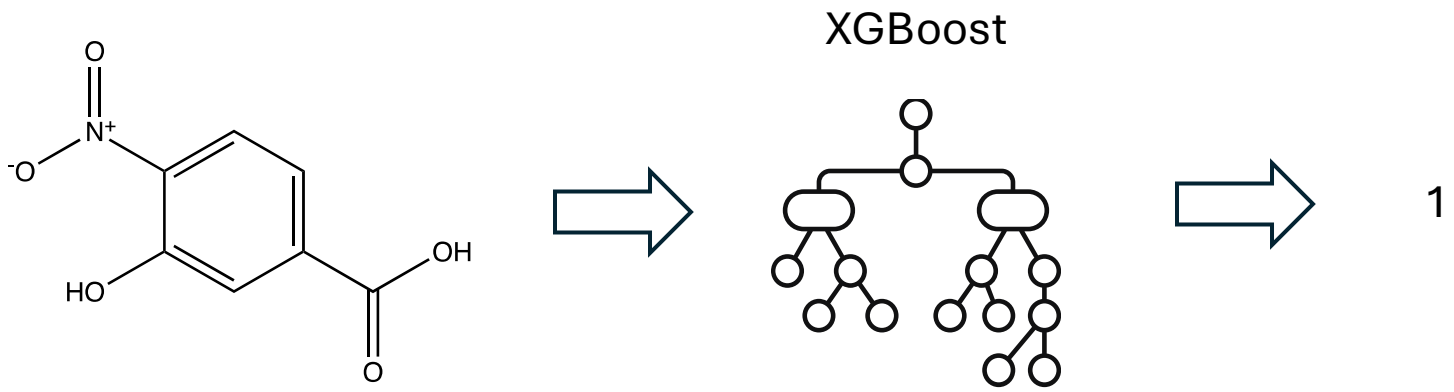
A faithful explainable AI (xAI) method should reflect the underlying mechanism of the model's predictive performance

= Model uses features of these atoms for its predictions

= xAI method highlights this group of atoms

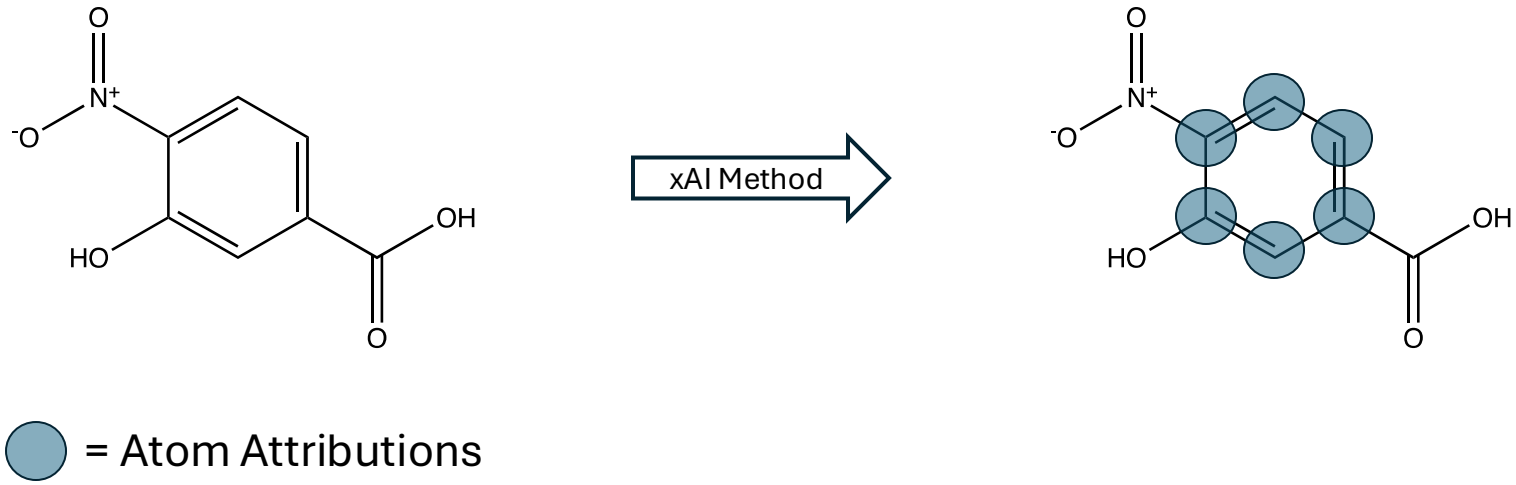
The Case Study

Very simple: Predicting the number of benzene rings in a molceule

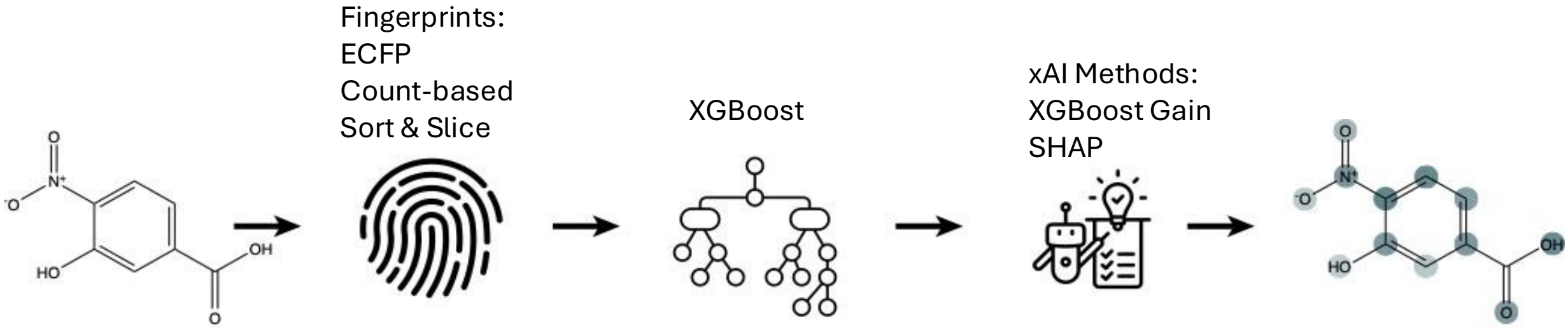


The Faithfulness Aspect

In a perfect world, a model with a R^2=1 should only use features that are associated with benzene



Experimental Setup

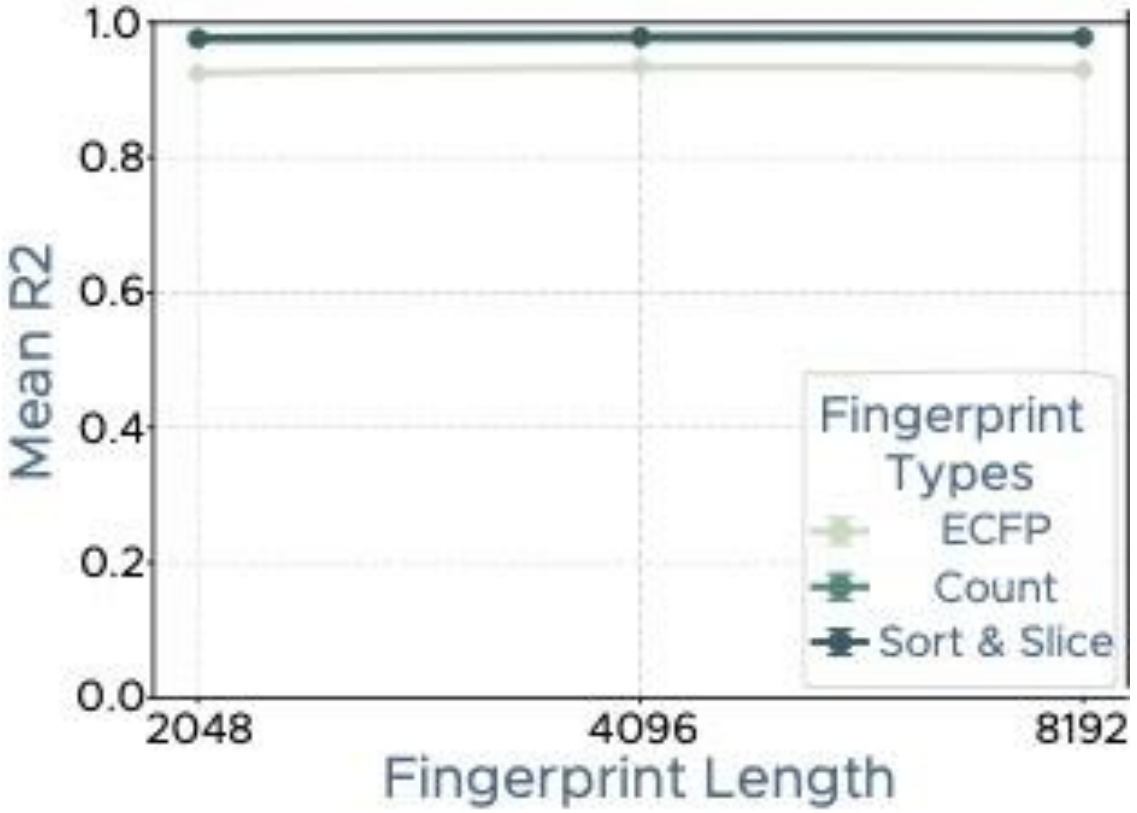


Fingerprint types
Hashed-based: ECFP, Count-based
Explicit (non-hashed): Sort & Slice

xAI methods
Global: XGBoost Gain
Local: SHAP

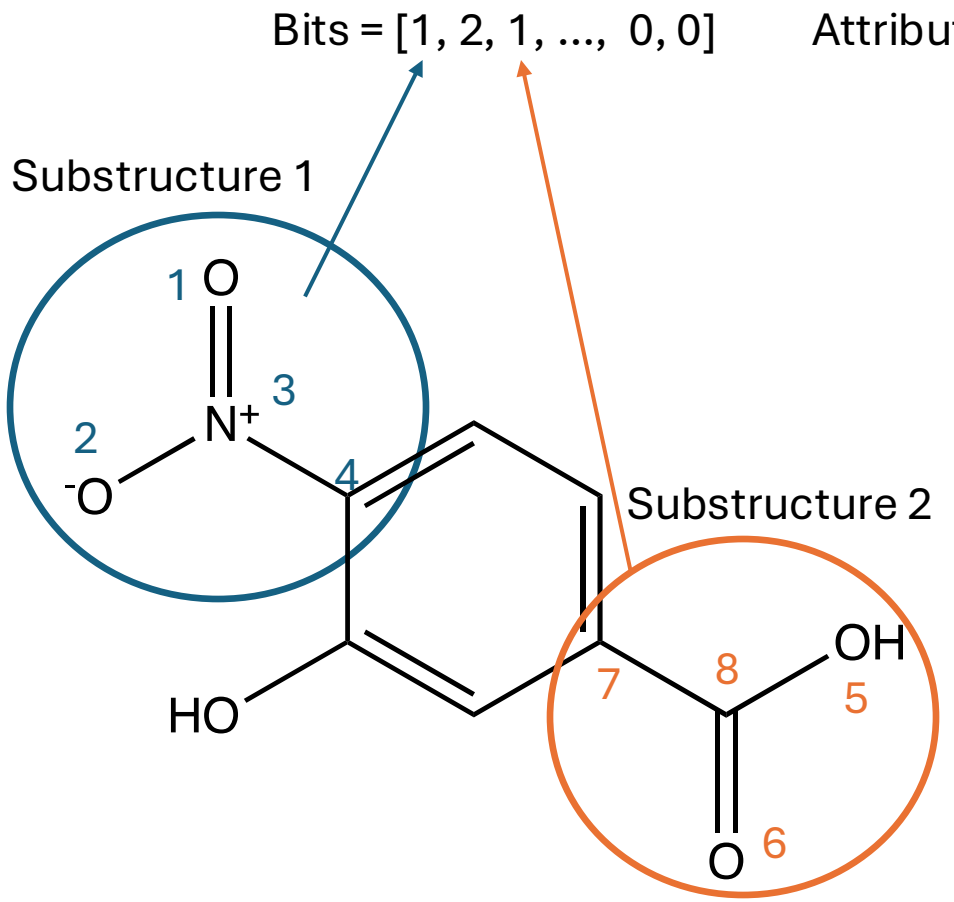
Results: Model Training

- Hyperparameter optimized on 5 random splits
- Result is the mean of the 5 test splits



Mapping back the attribution scores

Example: Bits with a counts and without bit collisions



Attributions atom 1 = $\frac{0.1}{1}$

Attributions atom 2 = $\frac{0.1}{2}$

Attributions atom 7 = $\frac{0.5}{2}$

Attributions atom 8 = $\frac{0.5}{2}$

All atoms of a given substructure will get the same attribution score

Faithfulness Metric

Model should learn the number of benzene rings, faithfulness metric should reflect that

$$a_i = \text{Atom Attributions}$$

$$F_{benzene} = \frac{\sum_{benzene} a_j}{\sum_{all} a_i}$$

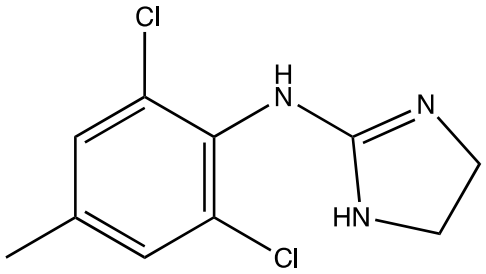
The higher the fraction is the better

Results: Benzene Fraction

Even with nearly perfect models, Benzene does not get the full attribution mass!

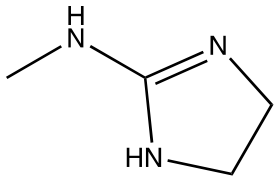
Challenge 1: Shortcut Learning

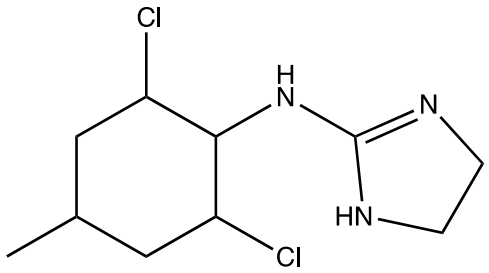
Example from Sort & Slice (2048 length)



0.24967189

This is the substructure with the highest importance:



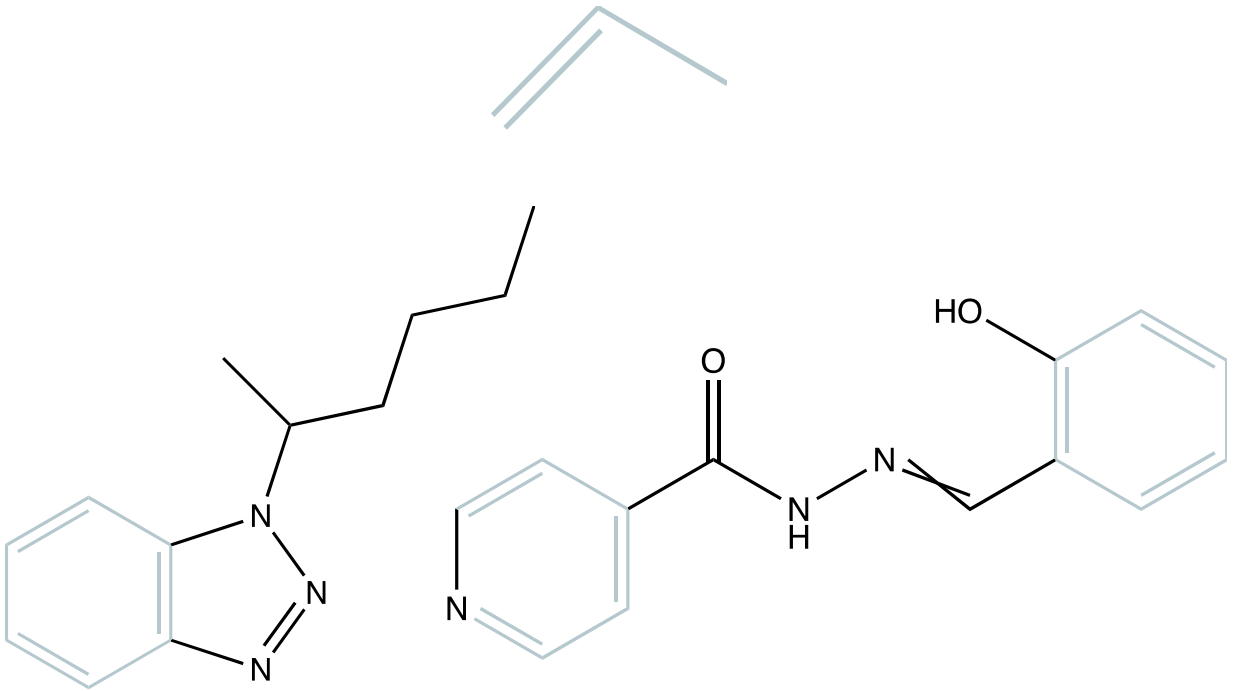


0.02065856

Often not easy to find and evaluate these shortcuts.
Faithfulness evaluation based on a given ground truth can be obscurred by shortcut learning.
Is the xAI method faithful in context of the shortcut or not?

Challenge 2: Fingerprints Granularity

Substructure



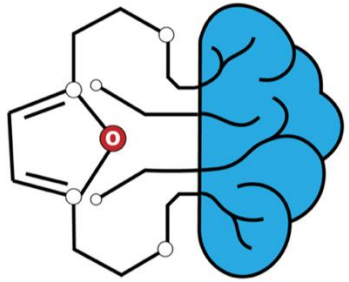
Diffusion of substructure importance to different chemical contexts

All the highlighted atoms get the same attributions. One does not now from which chemical context it gained the importance

Summary

- Effective evaluation requires more than a good performing model (shortcut learning, fingerprint granularity)
- Count-based fingerprints improve attribution focus compared to binary fingerprint (ECFP)
- No major differences between fingerprint sizes
- No major difference between hashed and non-hashed (Sort & Slice vs. Count-based)

Thanks



TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

ML MOLECULAR Machine Learning