Rostislav Fedorov | 12.10.2023

XAI Methods in Chemistry

How to explain a prediction?

What is being explored? Global property *vs.* local property.

What is the relation between the model and the interpretation? Intrinsic (ante-hoc) *vs.* extrinsic (post-hoc).

- Proposed Evaluation:[1]

Actionable, Complete, Correct, Domain Applicable, Fidelity/Faithful, Robust, Sparse/Succinct

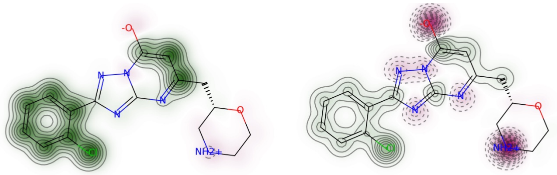
Subjective, as they depend on "complex human factors and application scenarios"

- Explanation technique

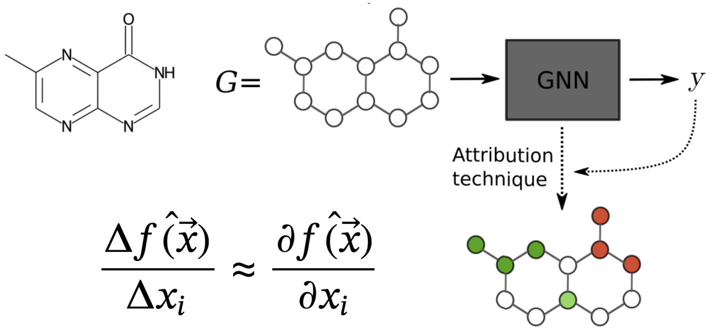
Self-Explaining Models

ML model[2]

Ground truth (Crippen's logP)



Attribution Methods[3]

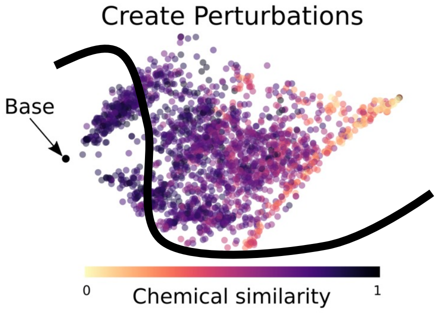


$$\frac{\Delta \hat{f}(\vec{x})}{\Delta x_i} \approx \frac{\partial \hat{f}(\vec{x})}{\partial x_i}$$

[1] *J. Chem. Theory Comput.* **2023**, *19*, 2149.
[2] *SciPost Chemistry* **2023**, *2*, 002. OA (CC BY license).
[3] Preprint DOI: 10.26434/chemrxiv-2022-v5p6m-v3. OA (CC BY license).

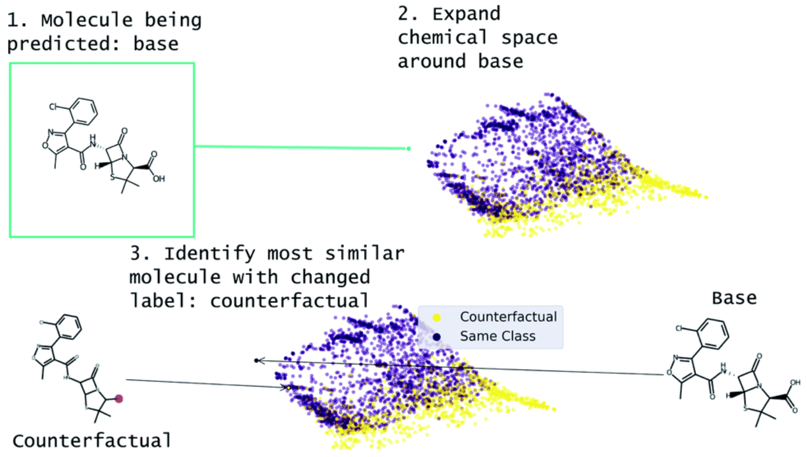
Surrogate Models[3]

$$\xi(\vec{x}) = \arg\min_{g\in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

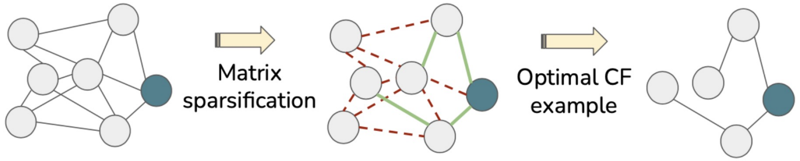


Counterfactual Explanations[4]

$$\text{minimize} \quad d(x, x')$$
$$\text{such that} \quad \hat{f}(x) \neq \hat{f}(x')$$


1. Molecule being predicted: base
2. Expand chemical space around base
3. Identify most similar molecule with changed label: counterfactual
Counterfactual
Same Class
Base
Counterfactual

CF-GNN Explainer[5]



[4] Chem. Sci. **2022**, *13*, 3697. OA (CC BY license).
[5] PMLR **2022**, *151*, 4499. Copyright 2022 by the author(s).