Explainability for molecular neural networks

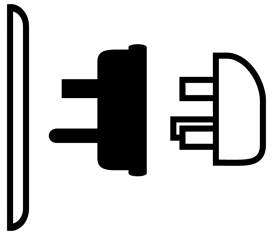
- **Floriane Montanari**

AIDD Summer School

May 2022

Many types of explainability approaches

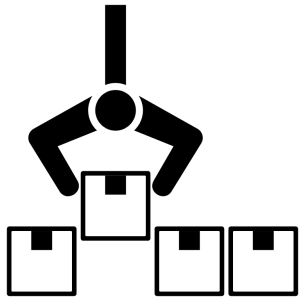
Model-agnostic methods



Once the model is built:

- Check for feature effects (PDP, ALE, feature importance, Shapley values, …)
- Approximate the model (global surrogate, LIME, anchors…)

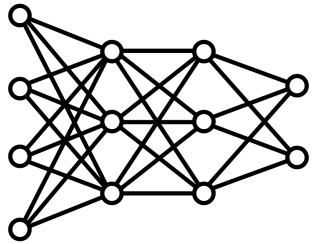
Example-based methods



Explain a trained model using examples:

- Counterfactual examples
- Adversarial examples
- Prototypes & critics
- Influential instances
- …

Neural networks-specific methods



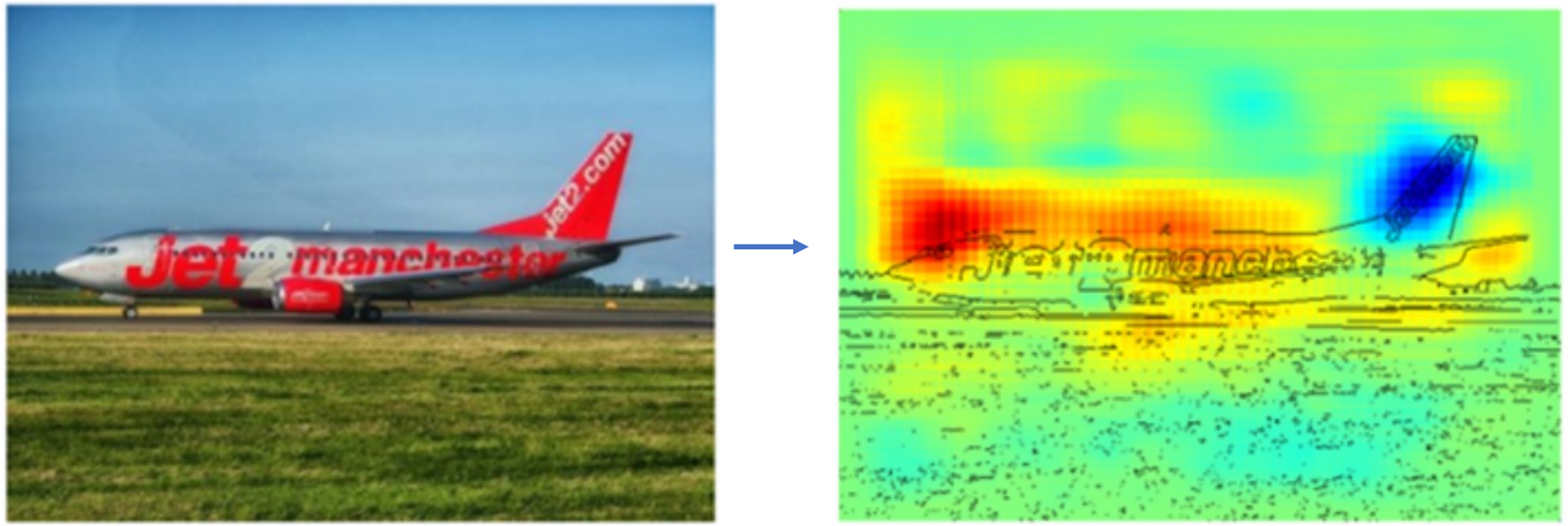
Specifically for image processing models. Visualize:

- Learned features and concepts
- Parts of input responsible for prediction (saliency methods)

Example of network-specific methods: explanations by saliency

<u>Intuition:</u>
Understand and measure input contribution to the output prediction.

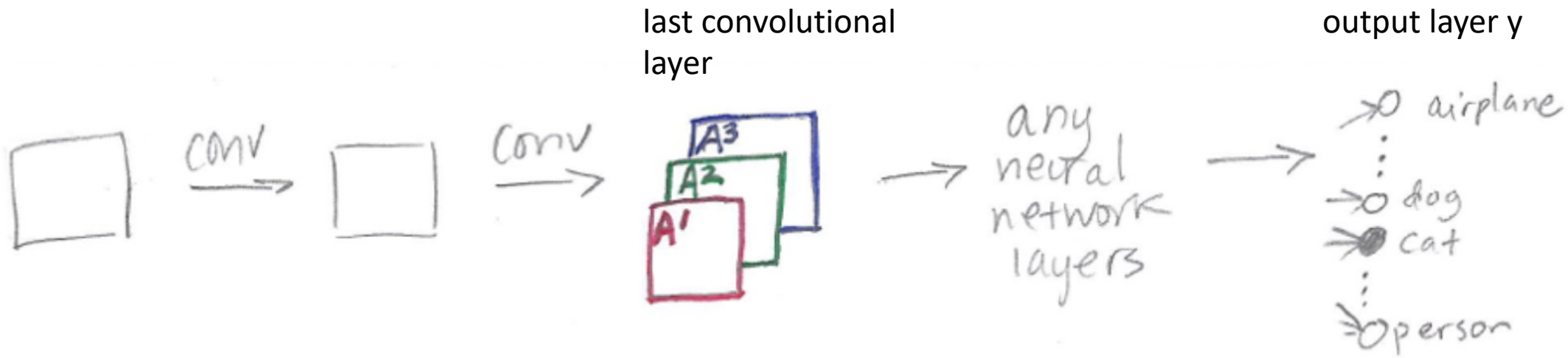
Occlusion maps
Grad-CAM
Integrated Gradients
Layer-wise relevance propagation
Deep Taylor decomposition
Spectral Relevance Analysis
Sensitivity analysis
Guided back-propagation
…



Input Image Heatmap

Bach et al., Plos ONE, 2017

How Grad-CAM works



1. For class *c* of interest, **compute gradient** of the output y_c with respect to the feature map activations A^k
2. For each map, compute an α_c^k value that is the **average of the gradients** across width and height
3. Return a weighted combination of the feature map activations A^k: $L^c = ReLU(\sum_k \alpha_k^c A^k)$
4. Scale up the heatmap to reach the initial image size (upsampling)

https://glassboxmedicine.com/2020/05/29/grad-cam-visual-explanations-from-deep-networks/

Interpretability for small molecules

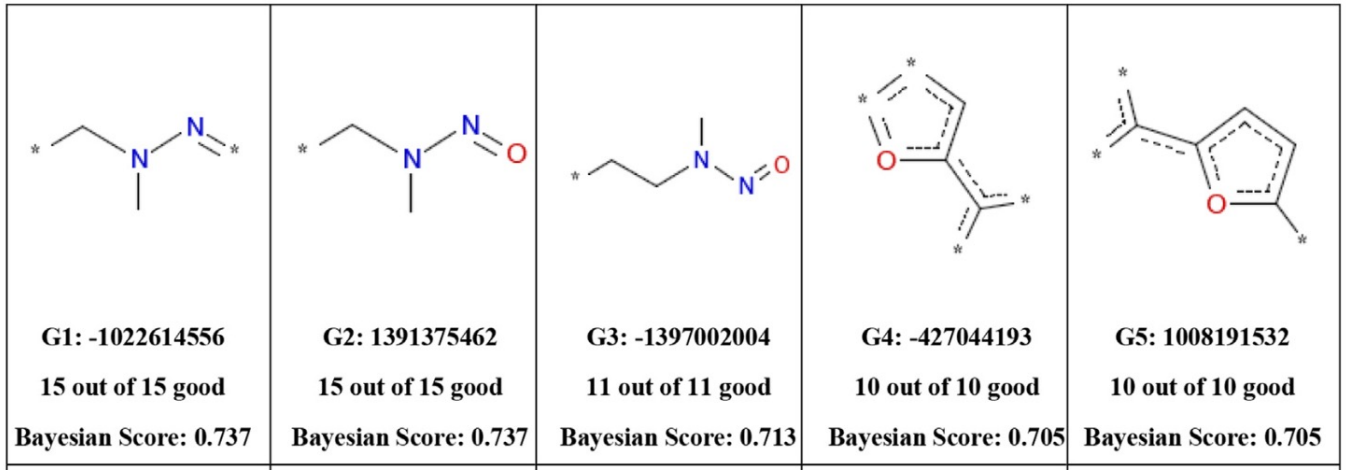
Aim

Which attributes of the molecule of interest are responsible for the prediction?

- Model debugging (is it learning what it is supposed to?)

- Insights for user on how to modify the structure to improve a desired property

Earlier attempts in the literature: Naïve Bayes on ECFP fingerprints



G1: -1022614556	G2: 1391375462	G3: -1397002004	G4: -427044193	G5: 1008191532
15 out of 15 good	15 out of 15 good	11 out of 11 good	10 out of 10 good	10 out of 10 good
Bayesian Score: 0.737	Bayesian Score: 0.737	Bayesian Score: 0.713	Bayesian Score: 0.705	Bayesian Score: 0.705

Global explanation of the model, based on training data bit frequencies.

Build a Naïve Bayes model using unfolded ECFP fingerprints

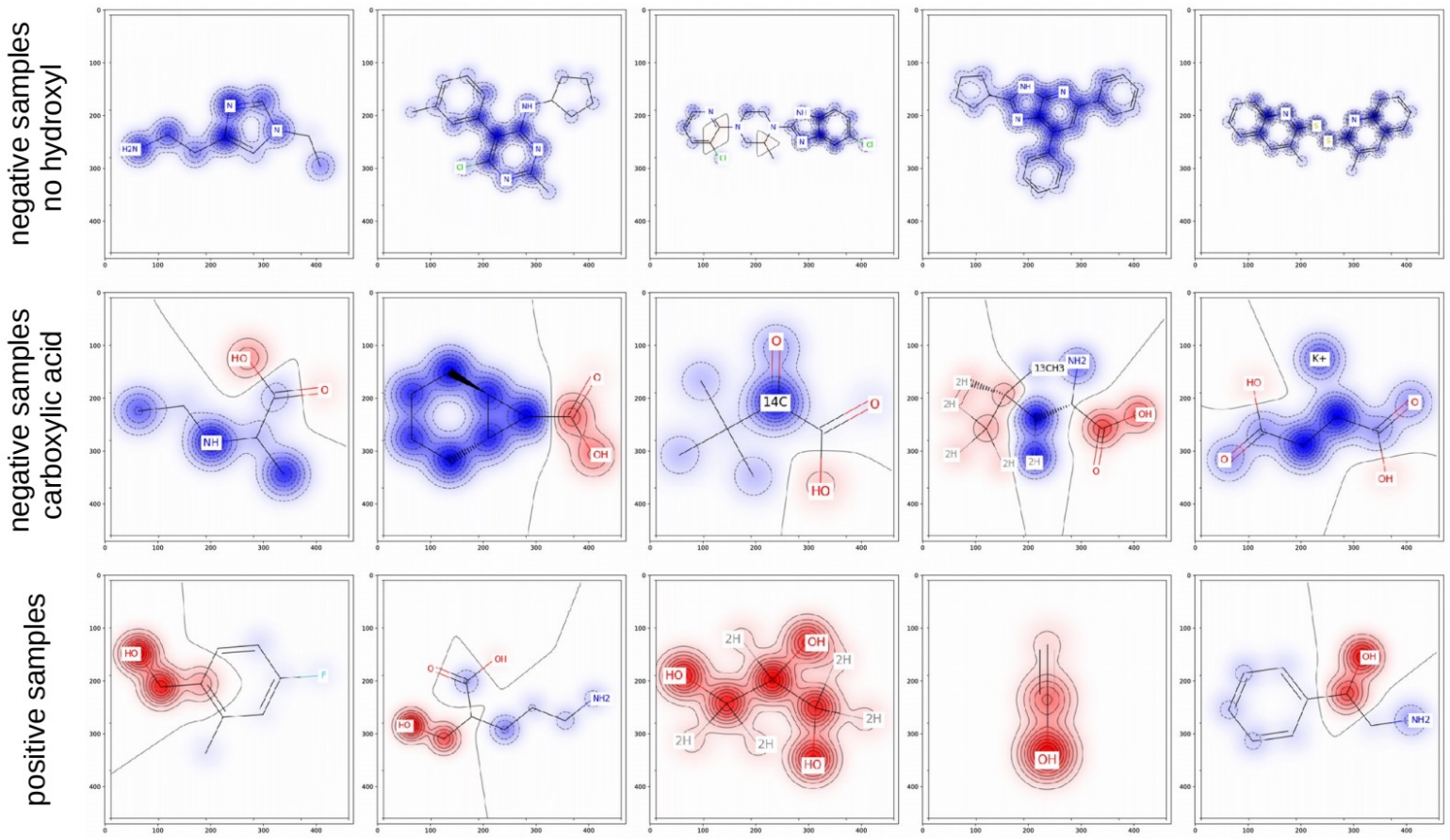
Retrieve the top N fingerprint bits that have the highest absolute value of Bayesian score

Bayesian score: how different is the observed ratio of occurrences in both classes compared to the expected if the feature was occurring randomly across both classes

Depict the fingerprint bits a molecular substructures

Zhang et al., 2016, Novel naïve Bayes classification models for predicting the carcinogenicity of chemicals

Earlier attempts in the literature: Fully-connected network based on ECFP fingerprints



Build a neural network using ECFP1 fingerprints on a toy dataset (task: alcohol group classification)

Apply Integrated Gradient to get back feature importance for the fingerprint bits

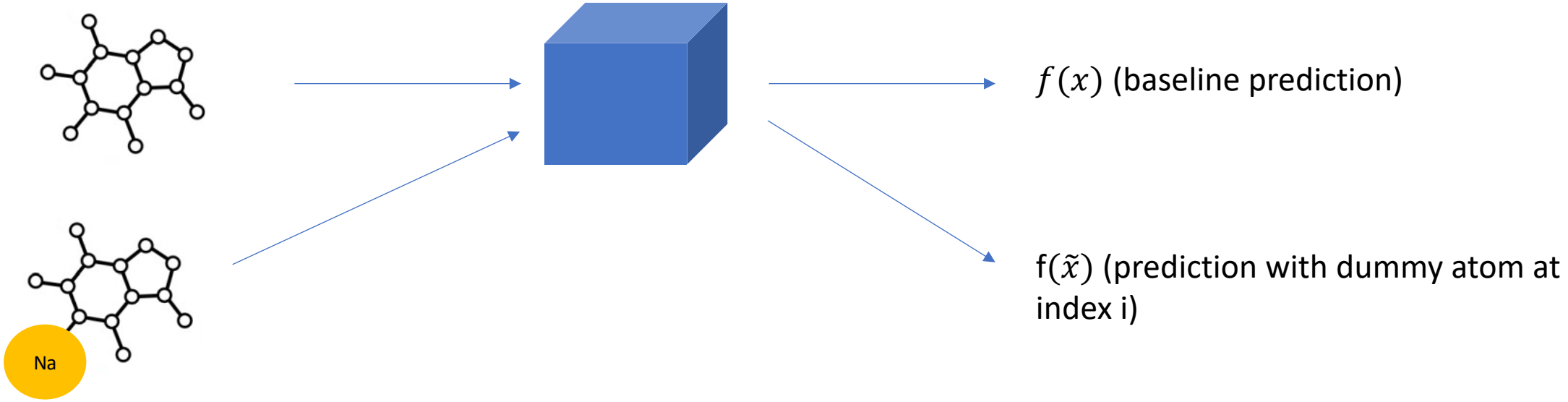
Get back to the underlying atoms: "Each fingerprint consists of multiple atoms and one atom is part of multiple fingerprints. Hence, we calculated the atom-wise attribution as the sum of the attributions of all fingerprints in which this atom is part of."

Preuer et al., 2019, Interpretable Deep Learning in Drug Discovery

“Sheridan” approach: dummy atom replacement

Input test molecules

Trained QSAR model



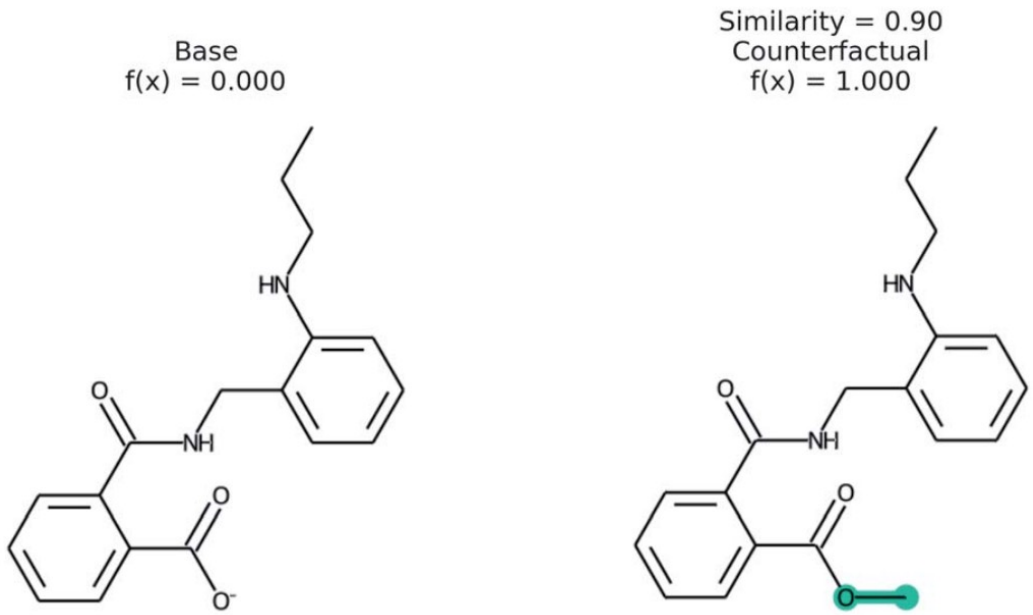
$f(x)$ (baseline prediction)

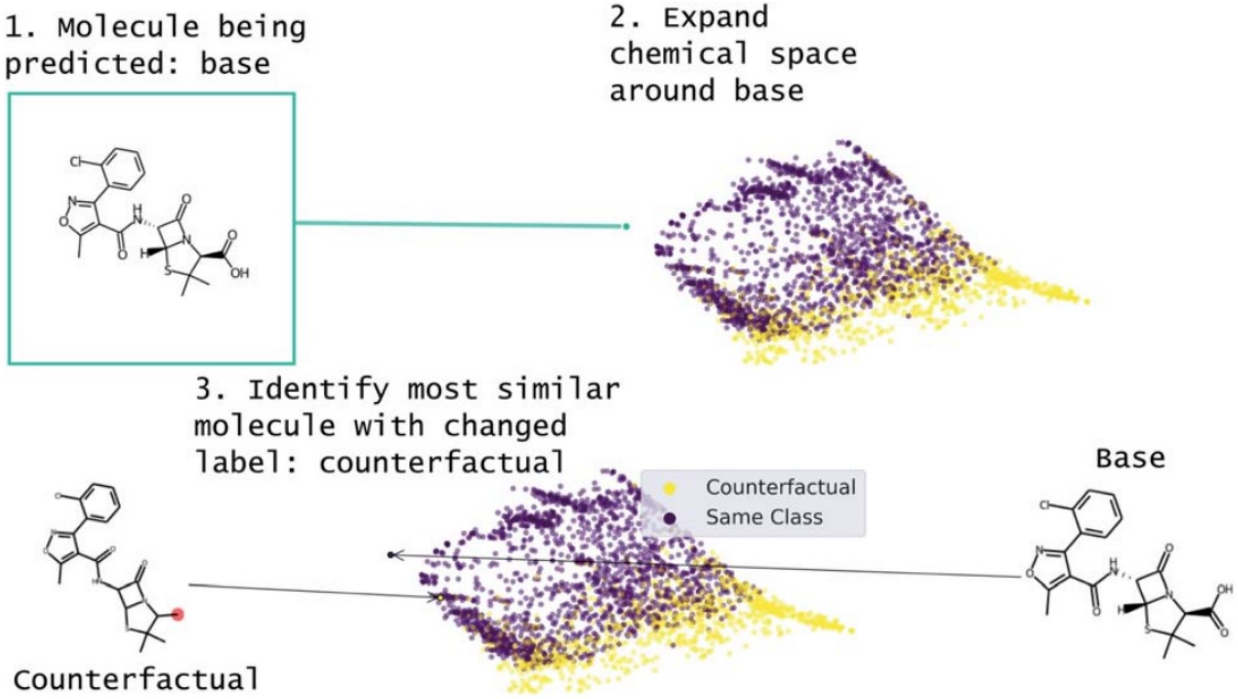
$f(\tilde{x})$ (prediction with dummy atom at index i)

$$Att_i = f(x) - f(\tilde{x})$$

Sheridan 2019, Interpretation of QSAR models by coloring atoms according to changes in predicted activity: how robust is it? JCIM

Example-based approache: counterfactual explanations

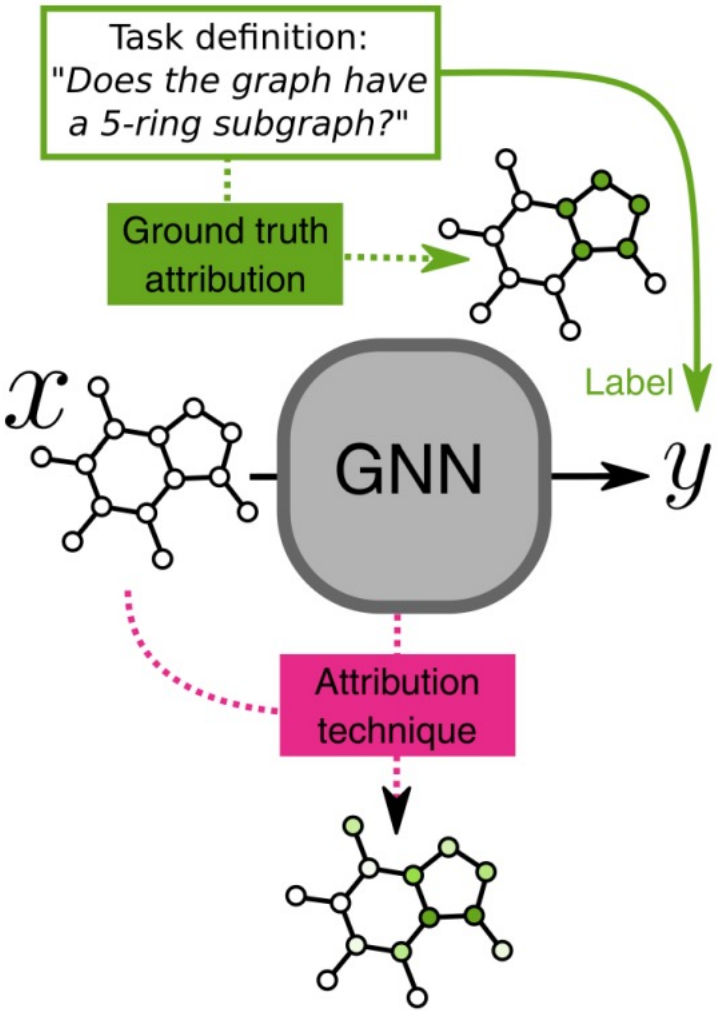




$$\begin{aligned} \text{minimize} \quad & \mathrm{d}(x, x') \\ \text{such that} \quad & \hat{f}(x) \neq \hat{f}(x') \end{aligned}$$

Wellawatte et al., 2022, Model agnostic generation of counterfactual explanations for molecules, Chemical Science

Evaluating attributions



In most chemical use cases, we do not have access to the ground truth of the explanation. We know the relationship structure – property but not the individual contribution of each group / atom to that property.

To evaluate XAI methods, one can create an artificial benchmark dataset where the atom contribution is obvious.

Example: Does the molecule has a 5-membered ring?
The model is a classification model (yes / no label) but the ground truth for an ideal explanation is straightforward to obtain.

Sanchez-Lengeling et al., 2020, Evaluating attributions for graph neural networks.

Additional benchmark datasets

DrugXAI Suite (Rao et al., 2021, ICML workshop on explainability)

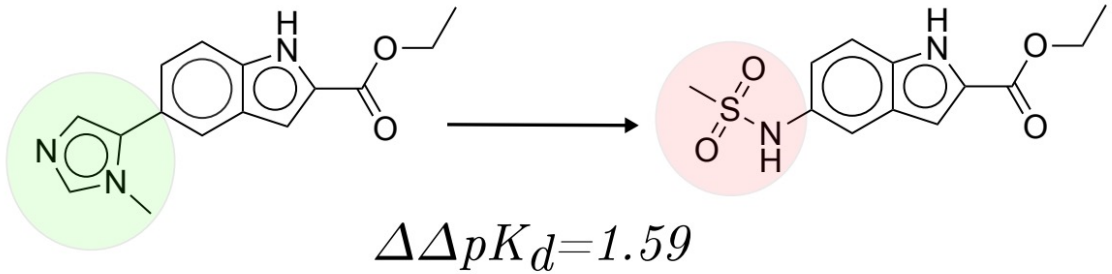
Table 1. Statisitics of the datasets

Tasks	Type	Dataset Name	Compounds	Train	Test	Subgraph ground truth
Single-rationale	Graph classification	3MR	3152	2521	631	Three-membered ring
	Graph classification	Benzene	12000	9600	2400	Benzene ring
Multiple-rationales	Graph classification	Mutagenicity	6506	5204	1302	Mutagenicity alerts
	Graph classification	Liver	587	469	118	Hepatotoxic alerts
Property cliff	Graph regression	hERG	6993	6483	510	Structural motifs
	Graph classification	CYP450	9122	9025	97	Structural motifs

similar to Sanchez-Lengeling

recognition of multiple rationales

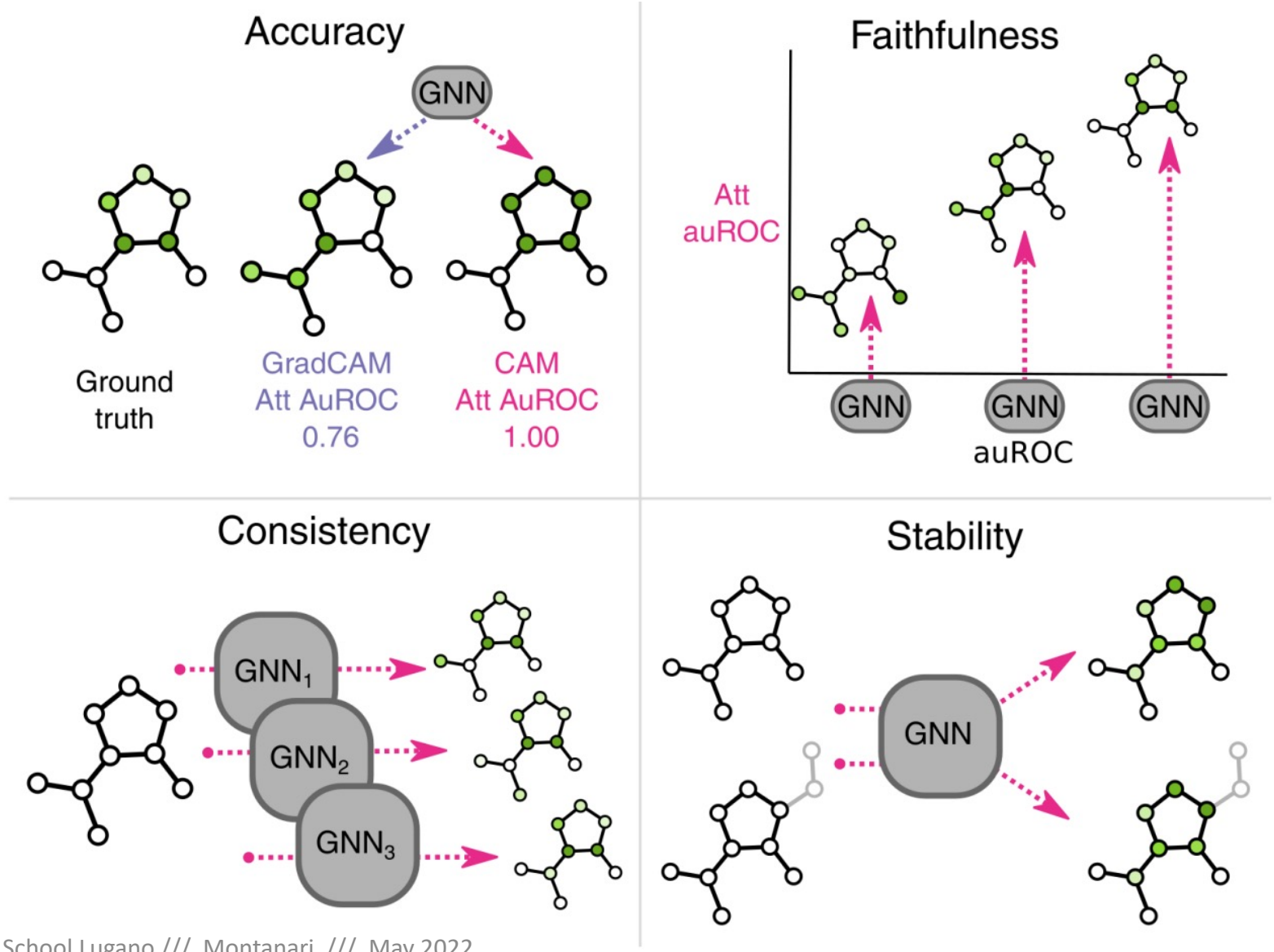
extremely difficult tasks, models tend to fail to properly predict the activity cliffs



Rao et al., 2021, Quantitative Evaluation of Explainable Graph Neural Networks for Molecular Property Prediction
Jimenez-Luna et al., 2021, Benchmarking molecular feature attribution methods with activity cliffs, JCIM

Evaluating attributions

Once the benchmark dataset is defined, one can think of possible metrics to evaluate the performance of an XAI method:



Sanchez-Lengeling et al., 2020, Evaluating attributions for graph neural networks.

Towards more interpretable graph convolutional models

Work mainly performed by Ryan Henderson

Background: graph convolutional model for physchem properties

Data

- 0.5M unique structures
- measured in 10 physchem assays (solubility, logD, melting point, etc.).

Model

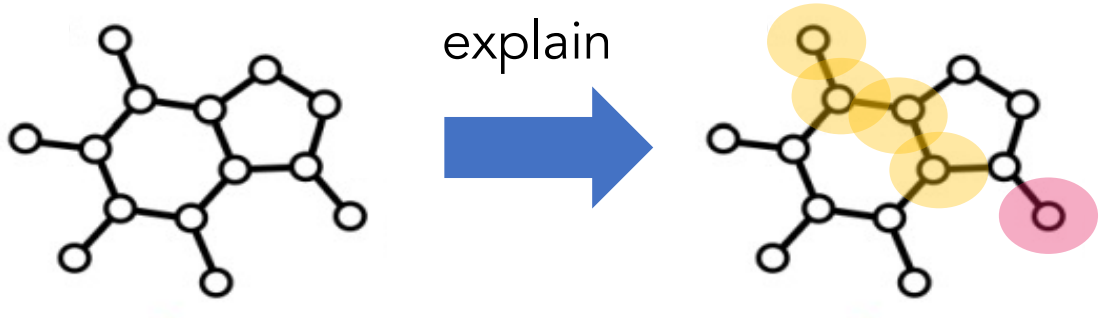
- multitask
- graph convolutional network with 2 convolutional layers
- input node representations computed with RDKit

Results

- Model currently in production, used by all pharma medicinal chemists

Montanari et al., 2021, Modeling physico-chemical ADMET endpoints with multitask graph convolutional networks

Background: interpretability for graph neural networks



Class activation maps
Input x Gradients
GradCAM
SmoothGrad
Integrated Gradients
Attention

Gradient-based methods

Graph-level tasks

	Benzene				Amine AND Ether AND Benzene				CrippenLogP			
	GCN	MPNN	GraphNets	GAT	GCN	MPNN	GraphNets	GAT	GCN	MPNN	GraphNets	GAT
Random Baseline	0.61	0.61	0.61	0.61	0.5	0.5	0.5	0.5	0.13	0.13	0.13	0.13
GradInput	0.72	0.54	0.54	0.56	0.52	0.53	0.55	0.41	0.12	0.09	0.13	0.1
SmoothGrad(GI)	0.71	0.54	0.54	0.53	0.51	0.55	0.59	0.38	0.15	0.11	0.15	0.11
GradCAM-last	0.74	0.72	0.66	0.66	0.54	0.74	0.55	0.46	0.04	0.33	0.24	0.07
GradCAM-all	0.75	0.68	0.84	0.62	0.54	0.62	0.7	0.44	0.05	0.27	0.27	0.09
IG	0.97	0.89	0.94	0.95	0.69	0.59	0.72	0.54	0.31	0.24	0.24	0.27
CAM	0.98	0.96	0.76	0.99	0.75	0.76	0.6	0.65	0.2	0.37	0.28	0.23
Attention Weights	--	--	--	0.51	--	--	--	0.51	--	--	--	-0.06

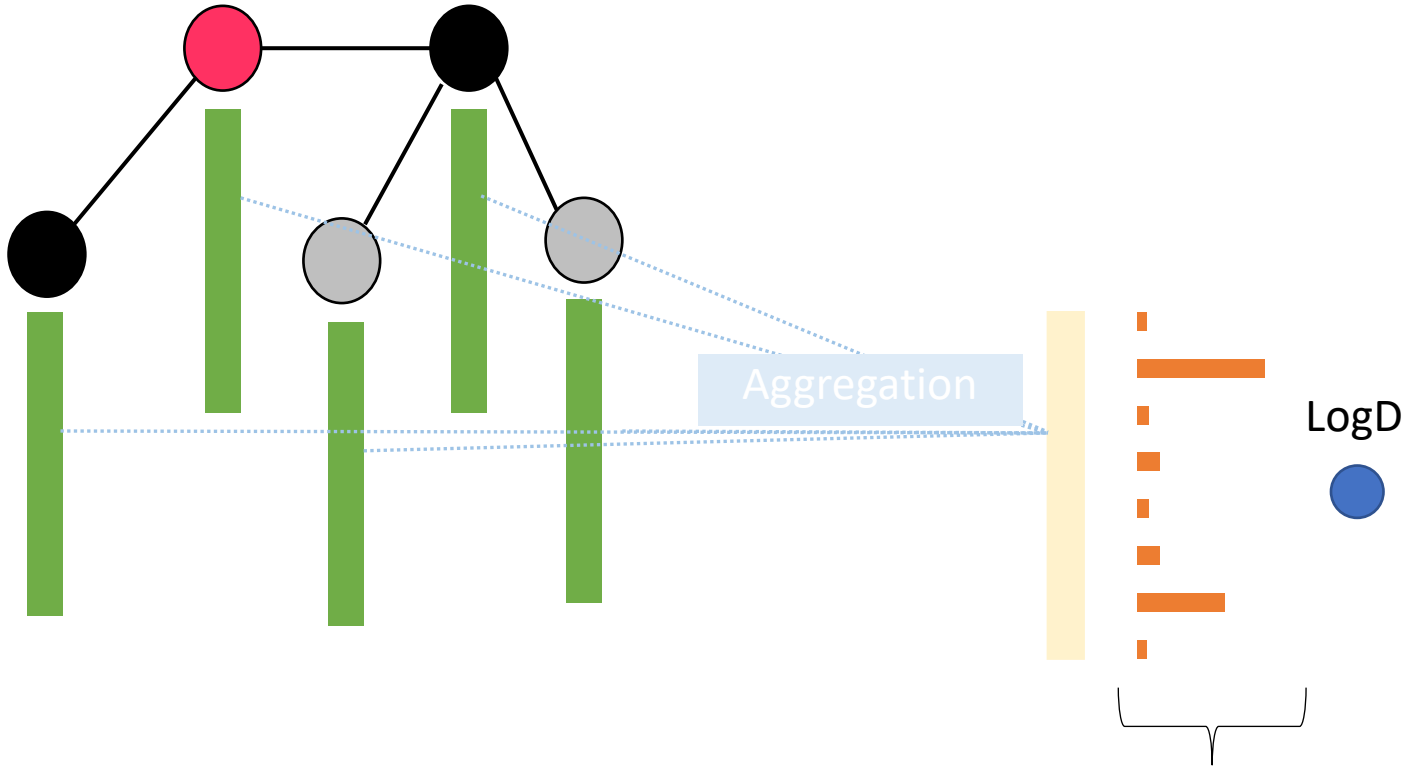
Sanchez-Lengeling et al., 2020, Evaluating attributions for graph neural networks.

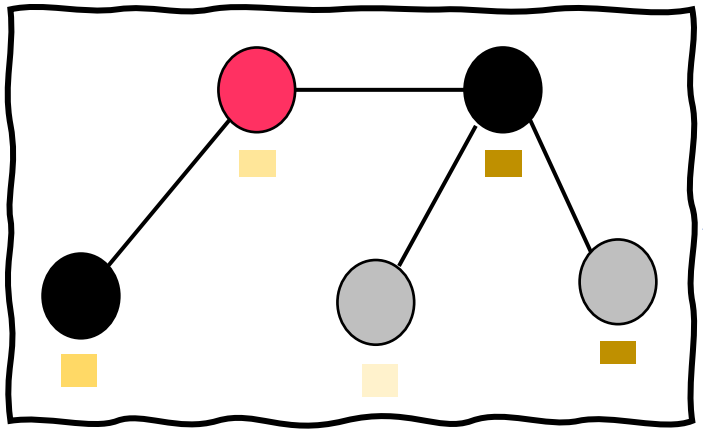
Class Activation Map (CAM) is a simple but powerful explainability method for GCN models

$$Att_v = w^T . act_v$$

Weights between last layer and output node of interest

Node *v* representation after the last convolution



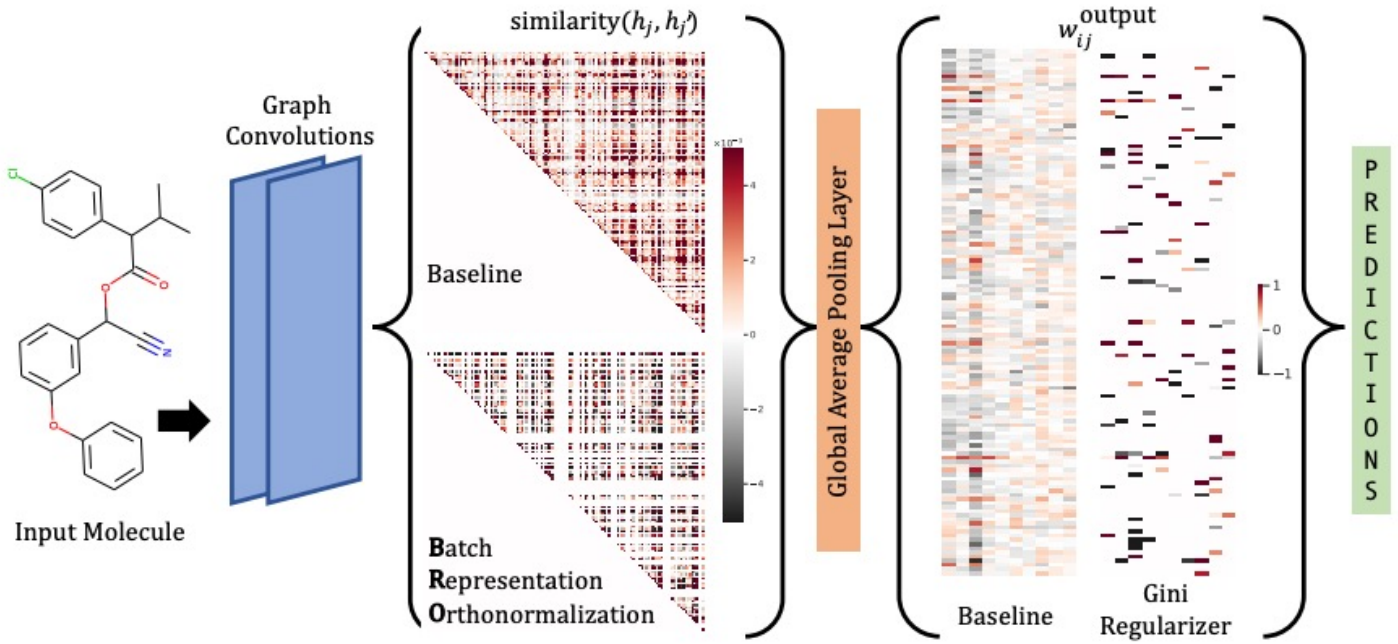


Attribution for logD mapped onto our demo graph

Adding constraints to the GCN architecture to improve interpretability

Reduce the number of non-zero weights on the output layer

Force each dimension of the atom representation to be independent from the others



Henderson et al., 2021, Improving molecular GCN explainability with orthonormalization and induced sparsity

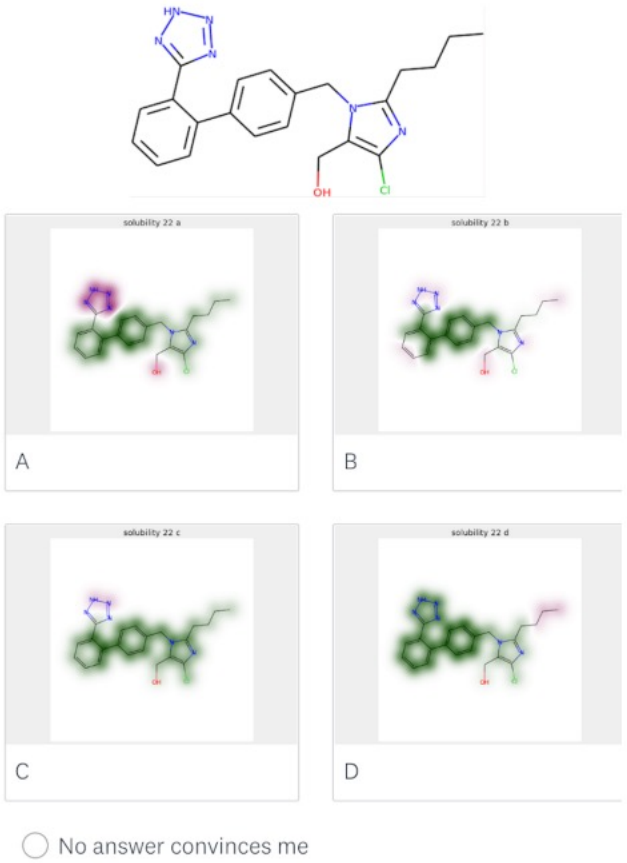
How does it work? (benchmark datasets)

Attribution performance:

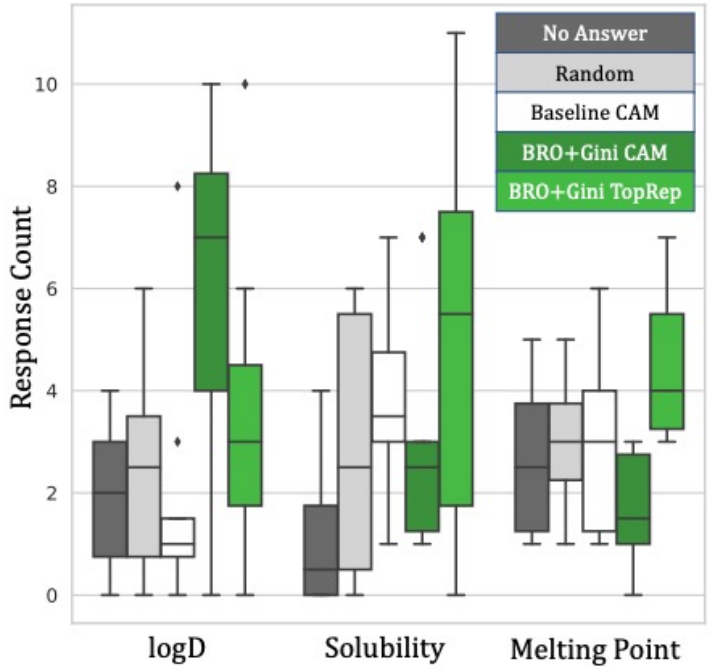
	Benzene task	Amine&Ether&Benzene task	CrippenLogP task
Baseline	0.99	0.62	0.27
BRO	0.99	0.62	0.27
Gini	0.99	**0.65**	**0.28**
BRO & Gini	0.99	**0.65**	**0.28**

On simple explainability benchmark datasets (finding benzene, finding amine and ether and benzene), and on different attribution methods, we see that the constraints can improve slightly the explainability performance.

How does it work? (real life example)



Survey sent to Bayer medicinal chemists, with answers provided by a baseline model, a randomly picked dimension in the atom representation, our BRO+Gini modified architecture, and the attributions obtained when looking only at the absolute highest weight in the output layer.



Lessons learned

For graph neural networks, providing explainability can be as simple as a product between output weights and last layer atom representation. This method is called CAM (Class Activation Map).

Adding some constraints to the architecture (sparsity, orthonormality) can help bring clearer explanations.

All architecture constraints can affect the overall performance of the model so it is a fine balance of performance vs explainability.

At Bayer, feedback from the medicinal chemists allowed us to discover some drawbacks of our solubility model. Work on architecture allowed us to improve the performance of the model.

Conclusions

Interpretability in machine learning is aimed at both the model developer and the model user.

In the specific case of QSAR model, one wants to see which parts of the input molecule affect most the predicted outcome.

True attributions are difficult to obtain so one need to use benchmarks or ask experts for feedback in order to evaluate new methods.

Other explainability methods like counterfactual explanations can also be explored, going beyond atom attributions.

Literature

- Interpretable Machine Learning: a guide for making black box models explainable, Book by Christoph Molnar, updated August 2021
- On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, Bach et al., *Plos ONE*, 2017
- Interpretable deep learning in drug discovery, Preuer et al., 2019, ArXiv
- Evaluating attributions for graph neural networks, Sanchez-Lengeling et al., 2020, NeurIPS
- Quantitative Evaluation of Explainable Graph Neural Networks for Molecular Property Prediction, Rao et al., 2021, ICML workshop on XAI
- Benchmarking Molecular Feature Attribution Methods with Activity Cliffs, Jimenez-Luna, JCIM, 2021
- Improving molecular GCN explainability with orthonormalization and induced sparsity, Henderson et al., 2021, ICML
- Novel naïve Bayes classification models for predicting the carcinogenicity of chemicals, Zhang et al., *Food Chem Toxicol*, 2016
- Interpretation of QSAR models by coloring atoms according to changes in predicted activity: how robust is it, Sheridan, *J Chem Inf Model*, 2019
- Model agnostic generation of counterfactual explanations for molecules, Wellawatte et al.., *Chemical Science*, 2022
- Drug discovery with explainable artificial intelligence, Jimenez-Luna et al., *Nature Machine Intelligence*, 2020
- Examples are not enough, learn to criticize! Criticism for interpretability, Kim et al., *NeurIPS*, 2016

Thank you!

- **Linlin Zhao, Henry Heberle**
- **Ryan Henderson, Marco Bertolini**
- **Okko Clevert**
- Sebastian Schmidt, Julian Heinrich
- Datavisyn, University Linz, Christina Hummer, Marc Streit
- **Funding: Bayer LSC Office**

floriane.montanari@bayer.com