


group of machine learning research



JAGIELLONIAN UNIVERSITY IN KRAKÓW

# B-XAIC Dataset:

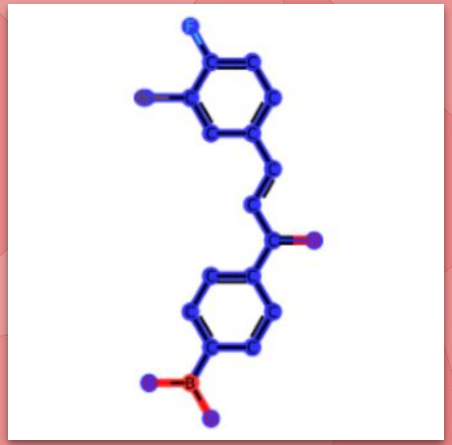
## Benchmarking Explainable AI for Graph Neural Networks Using Chemical Data

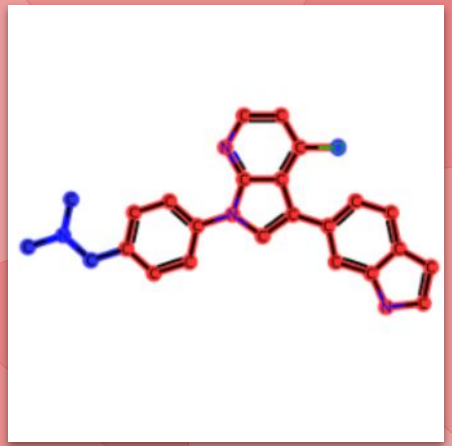
Magdalena Proszewska, Tomasz Danel, Antoni Antoszek, Dawid Rymarczyk

# Issues in XAI graph datasets

## Motivation for creating B-XAIC

- Real molecular data has no ground-truth explanations
- Existing benchmarks are small (MUTAG ~190 graphs) or synthetic
- Simple classification tasks
- Datasets assume at least one relevant substructure per graph.
- Most do not account for multiple valid explanatory patterns.





# B-XAIC content

## Molecules, tasks and explanations

- 50,000 real-world molecules from ChEMBL of average size ~34
- 7 binary classification tasks:
  - atom detection:
    - B
    - P
    - halogens (X)
  - indole ring detection
  - PAINS substructure detection
  - ring count > 4
  - maximum ring size > 6
- 5 splits
- Explanations
  - binary masks over atoms and bonds

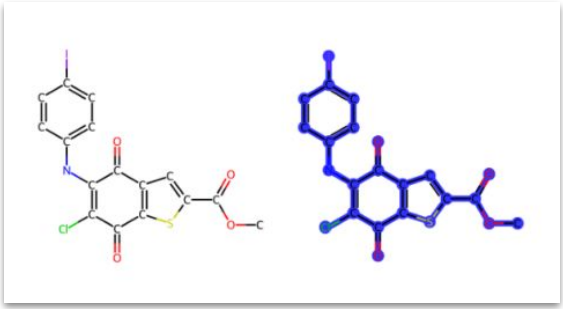
https://huggingface.co/datasets/mproszewska/B-XAIC

# Explanations in B-XAIC

## Two types of explanations and their evaluation

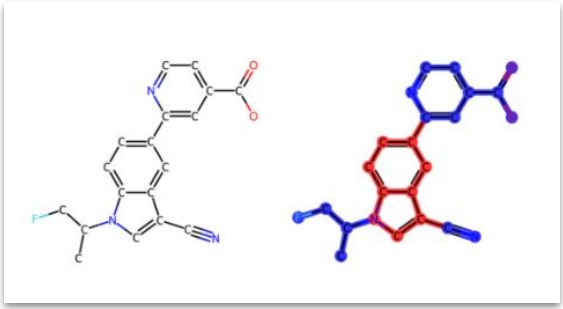
**Null explanations (NE)** – no nodes or edges are more important than others.

- Example: in the indole task, if the graph does not contain an indole substructure, then no specific substructure is considered relevant
- Evaluation using IQR outlier detection algorithm

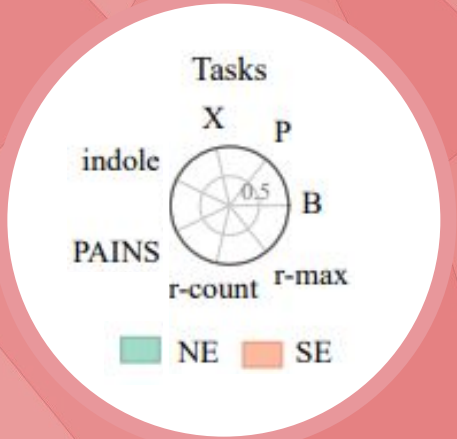


**Subgraph explanations (SE)** – only a part of the graph is relevant to the task.

- Example: in the indole task , if the graph does contain an indole substructure, that subgraph constitutes the explanation
- Evaluation using AUROC



Examples from the indole detection task

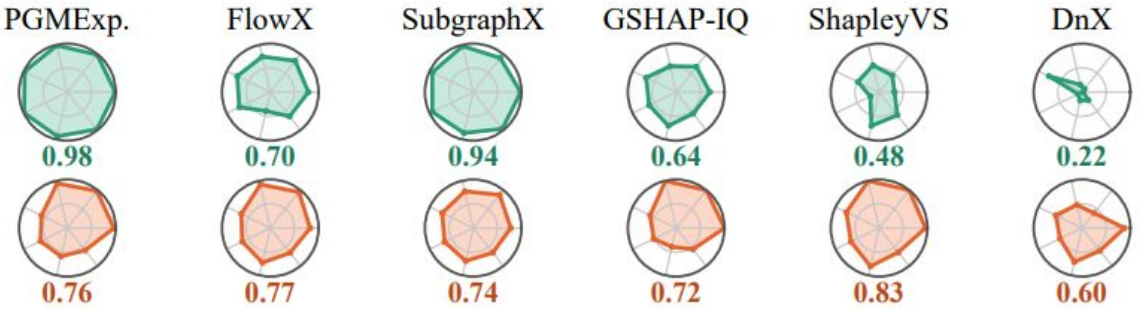



Visualization of evaluation scores

# Evaluation using B-XAIC

## Example application of B-XAIC and observations

Examples of evaluation results for 6/16 explainers applied on GIN models




We evaluated 16 explainers and observed that:

- sensitivity-based attribution struggles to represent feature absence,
- node and edge explanations can contradict each other within the same method,
- near-binary attributions may be insensitive to the evaluation threshold,
- no method is best on both NE and SE.

More explainers and models in the paper

# Applications of B-XAIC

## Intended use of our benchmark

- Benchmarking explainability methods for GNNs
- Assessing node- and edge-level explanations
- Developing interpretable machine learning models for graphs
- Research in explainable AI and cheminformatics

# Using B-XAIC

## Data and code

**HuggingFace**: https://huggingface.co/datasets/mproszewska/B-XAIC

**GitHub**: https://github.com/mproszewska/B-XAIC/

arXiv: https://arxiv.org/abs/2505.22252

```python
from bxaic import XAIMolecularDataset, TASKS

# Available molecular tasks
print(TASKS)
# ['B', 'P', 'X', 'indole', 'PAINS', 'rings-count', 'rings-max']

# Load the dataset for a specific task
dataset = XAIMolecularDataset("./data", "PAINS")
print(dataset[0])
# Data(x=[33, 11], edge_index=[2, 68], edge_attr=[68, 1], y=[1], expl_node_mask=[33], expl_edge_mask=[68])
```