GraphTrail: Translating GNN Predictions into Human-Interpretable Logical Rules

Burouj Armgaan*
Manthan Dalmia*
Sourav Medya
Sayan Ranu

What in the world...

What is GraphTrail?

A post-hoc, **global explainer** for message-passing GNNs that presents explanations as **logical rules** using **computation trees (CTrees)** as concepts.

And what is a global explainer?

A **local** explainer explains one graph at a time.

A **global** explainer explains an entire class of graphs at a time.

Why do we need global explainers?

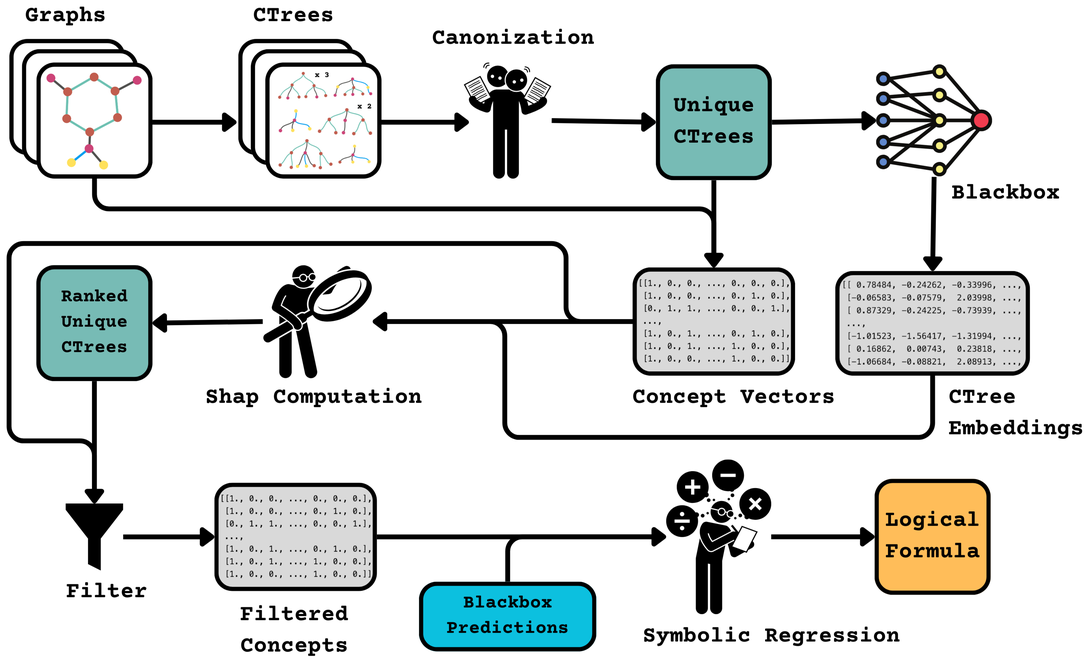
It takes a lot of **manual labor** to come up with a global understanding from local explanations.

If a prediction is based on the **absence of a part**, local explainers are stumped as the explanation is missing from the graph.

So... what's new about yours?

- No dependence on **third-party local explainers** for concepts
- Closer to a **GNN's true processing** as message-passing architectures operate on CTrees, not subgraphs
- Does not use **pseudo-concepts** like synthetic graphs and vectors. **Mines concepts** from the dataset
- Uses **symbolic regression** to identify the logic in which the GNN combines the concepts.

How do you do it?



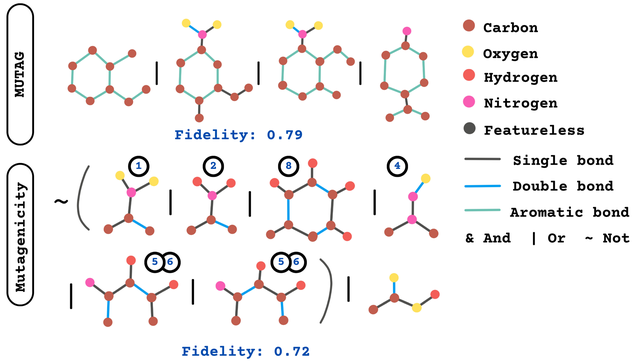
Any catches I should know of?

- Boolean logic **cannot count** concepts.
- Rooted-tree isomorphism limits GraphTrail to

How are the numbers?

	BAMultiShapes	MUTAG	Mutagenicity	NCI1
GLG	0.48 ± 0.02	0.74 ± 0.10	0.62 ± 0.03	0.57 ± 0.04

Show me some of your explanations



The encircled numbers in Mutagenicity are **toxicophore** identifiers. Mutagenicity's rule is for the non-mutagenic class.

Toxicophore name	Aromatic Nitro	Aromatic Amine	Three-membered heterocycle	Nitroso	Unsubstituted heteroatom-bonded heteroatom	Azo-type	Aliphatic Halide	Polycyclic Aromatic System
Substructure representation								
Identifier	1	2	3	4	5	6	7	8

According to the dataset authors [1], these eight **toxicophores** in the Mutagenicity dataset identify 75% of all mutagens.

References

[1] Jeroen Kazius, Ross McGuire, and Roberta Bursi. Derivation and validation of toxicophores for mutagenicity prediction.