Hierarchical Explainable Latent Pooling discovers hierarchical concepts in GNNs

[1] UNIVERSITY OF CAMBRIDGE

Jonas Jürß[1], Lucie Charlotte Magister[1], Pietro Barbiero[1], Pietro Liò[1], Nikola Simidjievski[1]

correspondence to: jj570@cl.cam.ac.uk

[2] Università della Svizzera italiana

Motivation

- Most **real-world graphs contain natural hierarchies** (e.g. functional groups in molecules/cliques in social networks/…)
- **Existing** interpretable GNNs **focus on the last layer** and do not give insights how concepts compose between layers

Method →

1. Apply GNN layers
2. Perform Clustering
3. Merge connected components of nodes in the same cluster
4. Repeat

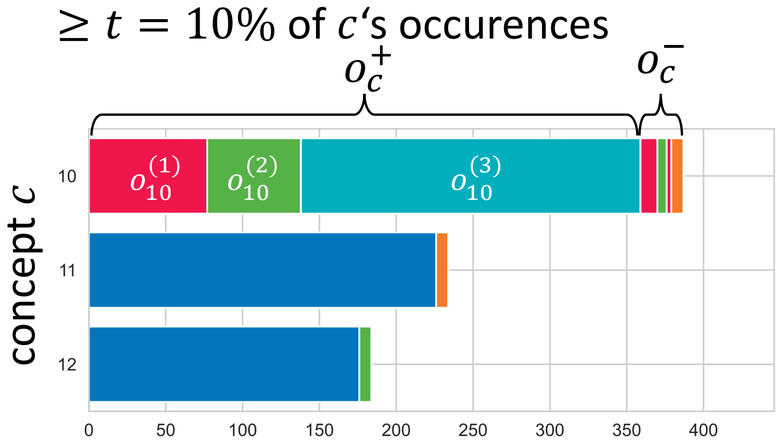
Concepts

... in this context are **components in a graph**, where mainly the presence of the **component is relevant for the final prediction, rather than each individual node** (e.g. functional groups for molecular property prediction)

Evaluation Metric:

Concept Conformity

components $o_c^{(j)}$ that make up $\geq t = 10\%$ of c's occurences



$$\text{conf}(c) = \frac{o_c^+}{o_c^+ + o_c^-}$$

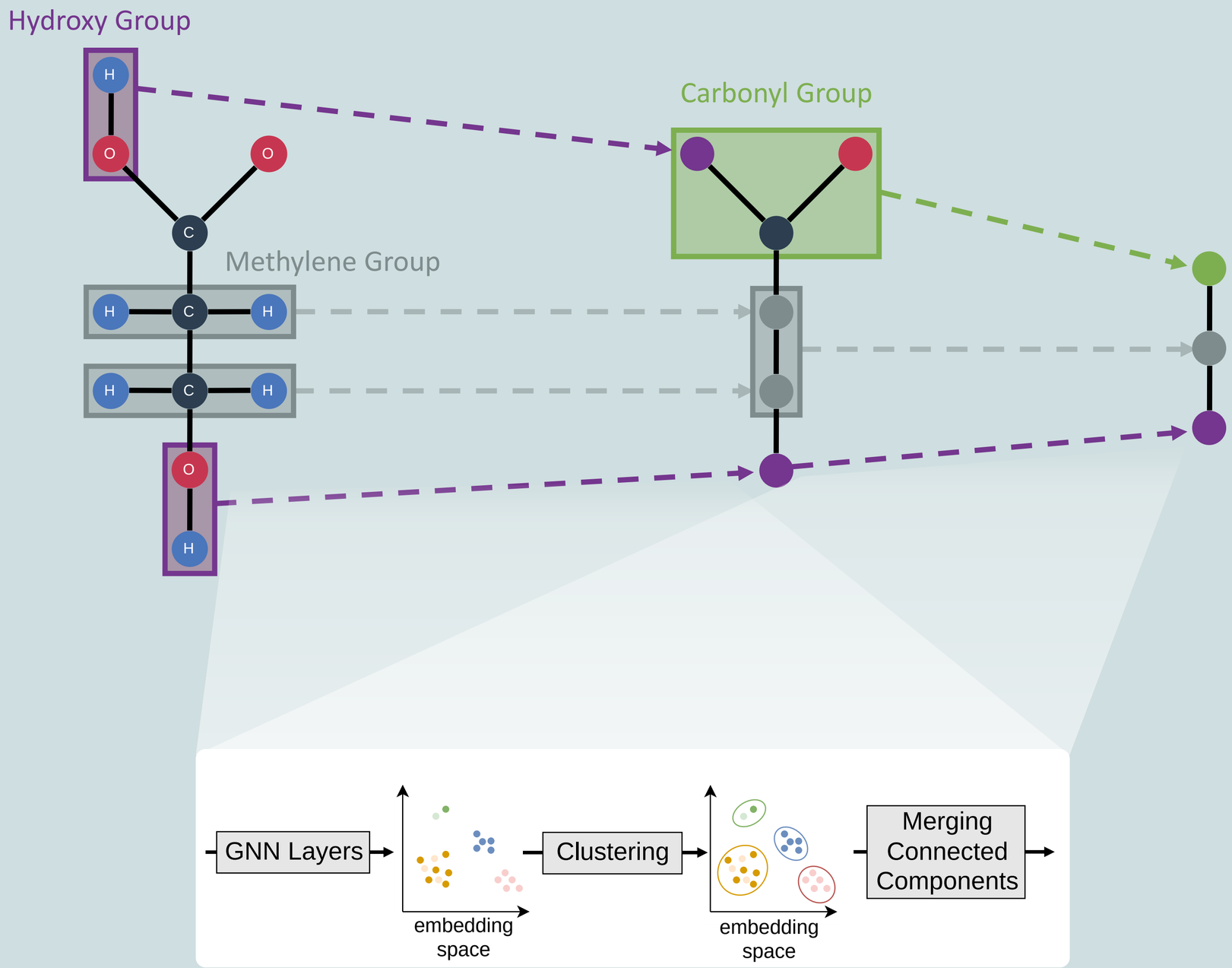
The model's conformity:
average $\text{conf}(c)$ over all concepts c

More Features

- More than 1-WL expressive
- First non-spectral method that can learn to pool arbitrary connected components

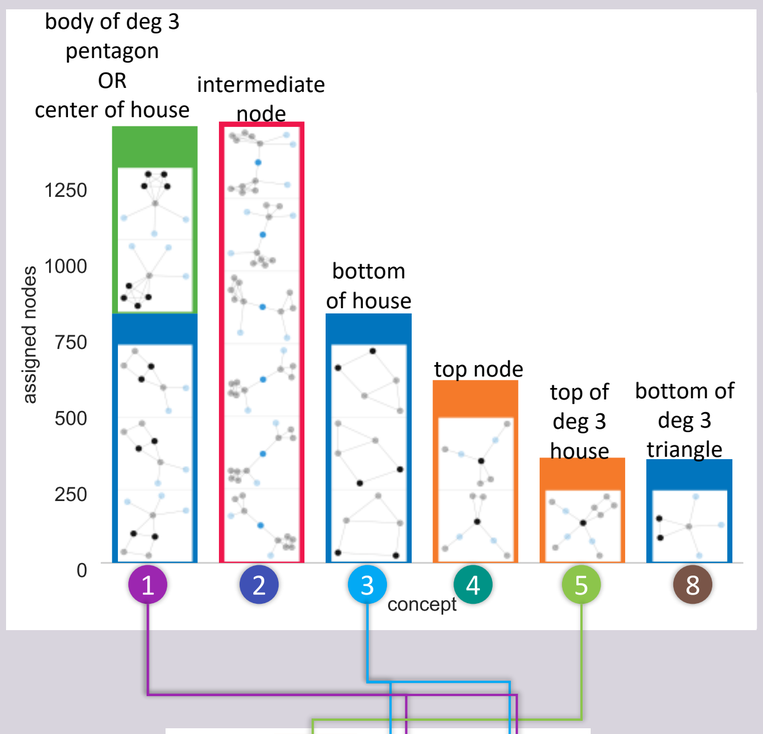
Paper

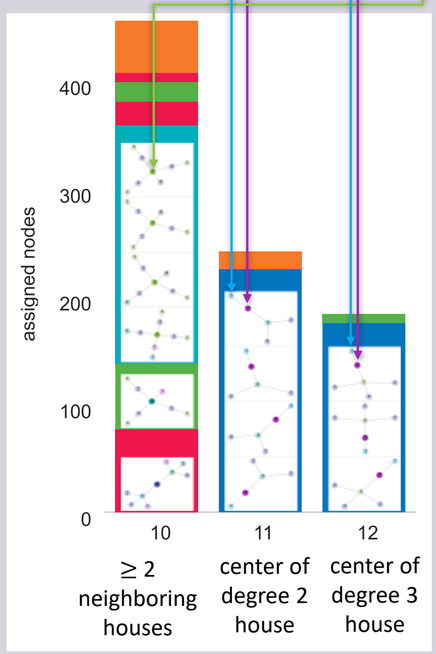
Code



Results

Discovering Hierarchies





Quantitative Results

TLDR; HELP performs on-par in terms of accuracy but yields cleaner & more detailed concept explanations

	HELP (Ours)			DiffPool		
	Acc.	Comp.	Conf.	Acc.	Comp.	Conf.
Synth. Hier.	99.9±0.2	**100.0**±0.0	**99.8**±0.4	**100.0**±0.0	27.0±0.0	0.0±0.0
Synth. Exp.	**100.0**±0.0	52.3±0.0	0.0±0.0	53.5±0.0	**53.5**±0.0	0.0±0.0
Mutag.	77.0±2.3	73.7±2.7	**83.6**±0.3	78.7±0.6	53.6±0.0	42.9±0.0
BBBP	85.0±1.6	80.8±1.4	**84.8**±1.4	82.0±5.6	77.1±0.4	0.0±0.0
REDDIT-BIN[4]	88.7±2.2	infeas.	**96.2**±0.4	**93.9**±0.7	infeas.	93.0±2.6

	ASAP			GCN	(+ GCExpl.)	
	Acc.	Comp.	Conf.	Acc.	Comp.	Conf.
Synth. Hier.	96.9±4.8	n/a	n/a	**100.0**±0.0	**100.0**±0.0	16.8±1.8
Synth. Exp.	93.9±0.2	n/a	n/a	53.5±0.0	**53.5**±0.0	**74.4**±0.0
Mutag.	76.2±1.7	n/a	n/a	**80.5**±0.7	**77.5**±2.4	16.5±10.1
BBBP	**85.2**±1.5	n/a	n/a	84.9±3.1	**86.0**±1.6	5.8±6.0
REDDIT-BIN[4]	infeas.	n/a	n/a	89.1±0.9	infeas.	infeas.

Real-World Examples of Discovered Concepts