

IOWA STATE UNIVERSITY

Department of Computer Science

Molecular Representation Learning via Heterogeneous Motif Graph Neural Networks

Zhaoning Yu and Hongyang Gao

Outline

- **What are motifs?**
- Why motifs?
- How to effectively use motifs

Motifs-based Molecular Representation Learning

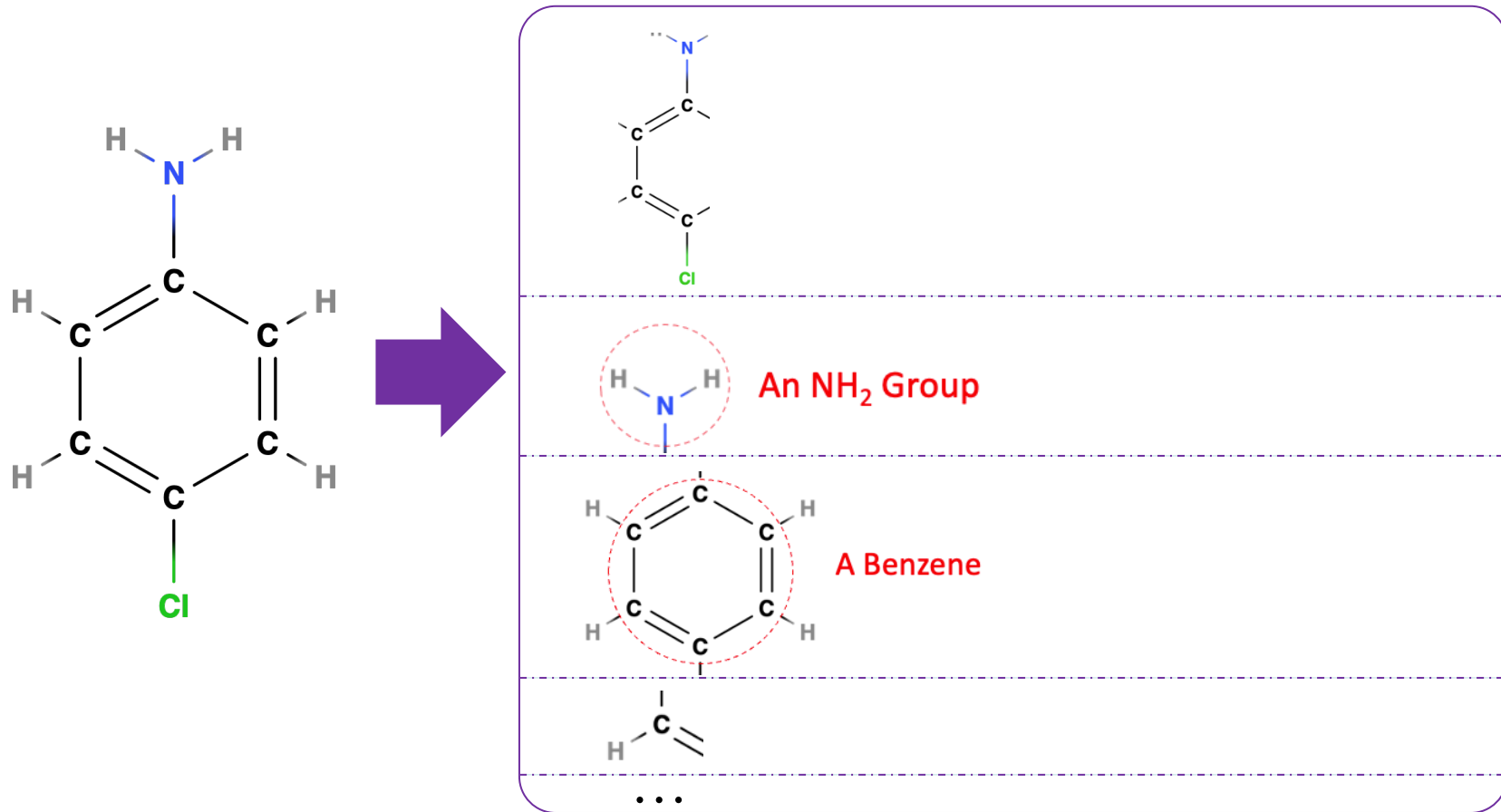
What are motifs?

- Motifs are **recurrent** and **statistically significant** subgraphs or patterns in a graph dataset.

Motifs-based Molecular Representation Learning

What are motifs?

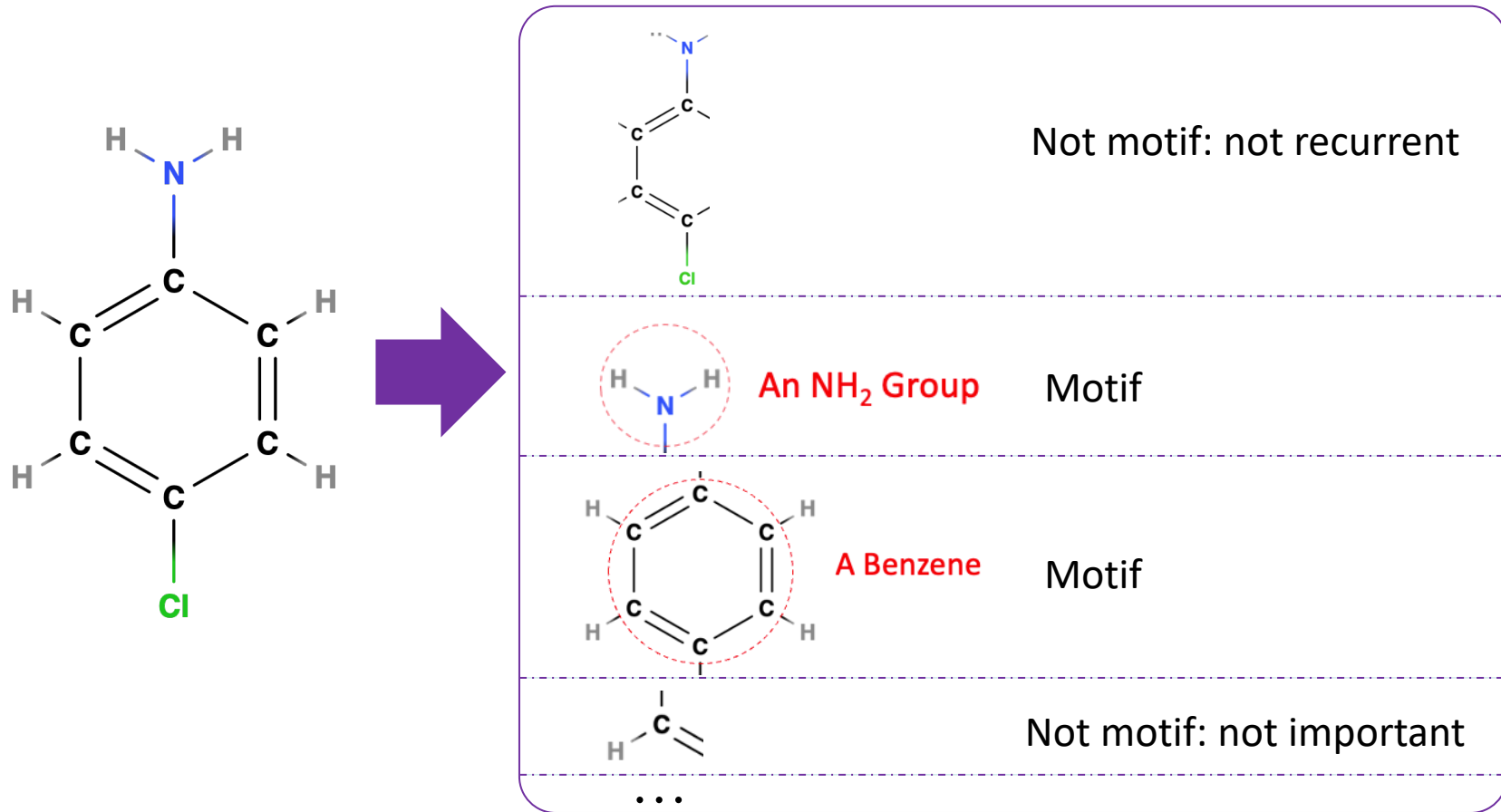
- Motifs are **recurrent** and **statistically significant** subgraphs or patterns in a graph dataset.



Motifs-based Molecular Representation Learning

What are motifs?

- Motifs are **recurrent** and **statistically significant** *subgraphs* or *patterns* in a graph dataset.



Motifs-based Molecular Representation Learning

Outline

- What are motifs?
- **Why motifs?**
- How to effectively use motifs

Motifs-based Molecular Representation Learning

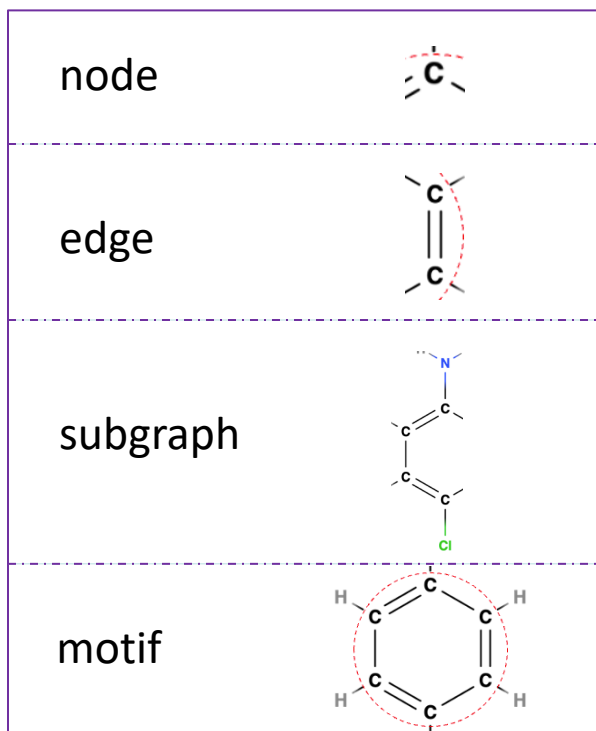
Why motifs?

- Motifs have been extensively studied in many fields
 - Biochemistry, ecology, neurobiology, and engineering
 - Proved to be important.
- Can capture structural information

Motifs-based Molecular Representation Learning

Why motifs?

- Motifs have been extensively studied in many fields
 - Biochemistry, ecology, neurobiology, and engineering
 - Proved to be important.



NLP

character	<i>a, b, c</i>
adjacent characters	<i>ab, bc</i>
span	<i>otif, mo</i>
word	<i>motif</i>

Motifs-based Molecular Representation Learning

Outline

- What are motifs?
- Why motifs?
- **How to effectively use motifs**

How to effectively use motifs

- A straightforward way: collect motif information in a molecule
 - Bag of motifs
 - Cannot capture interactions among motifs

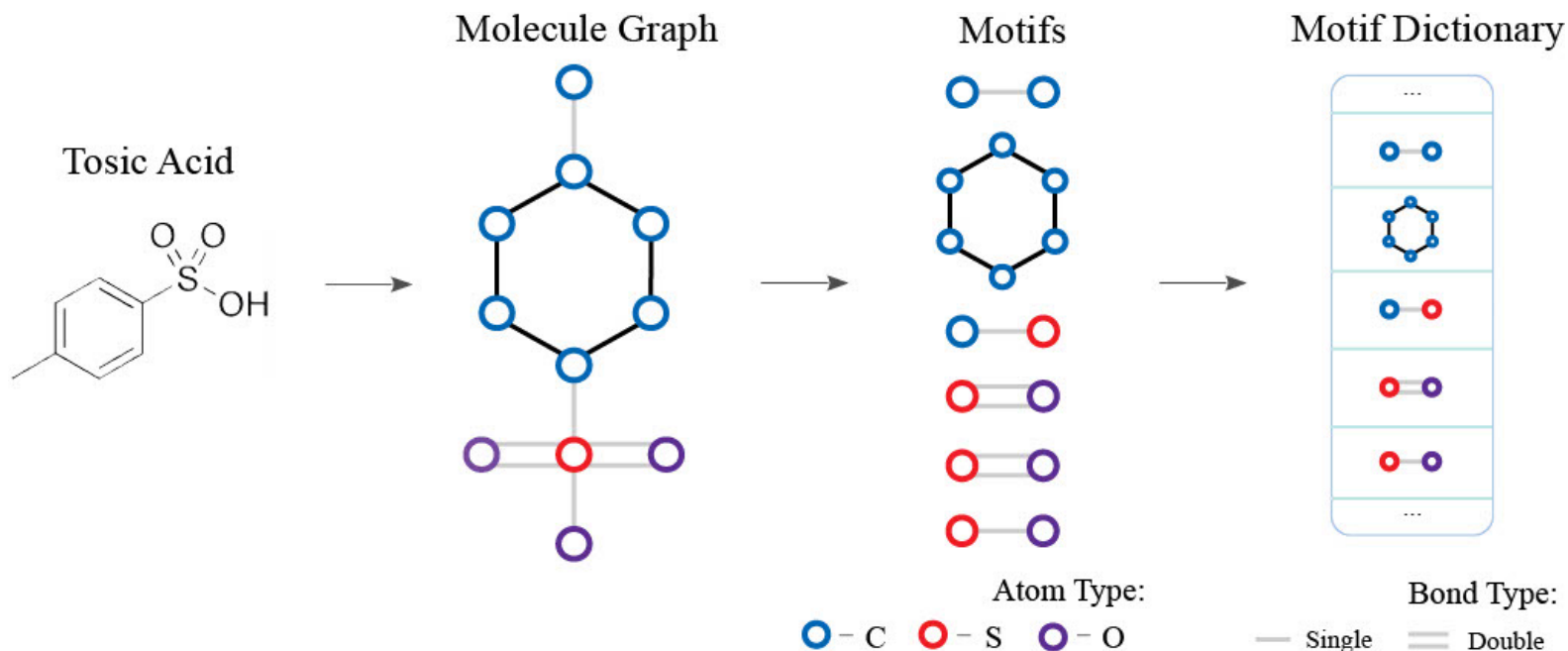
If a molecule contains both positive and negative motifs, what would be its property?

Need to consider motifs relationships

Our Method

- Build a motif vocabulary
- Construct a Heterogenous Motif Graph
- Learning based on this Heterogenous Motif Graph

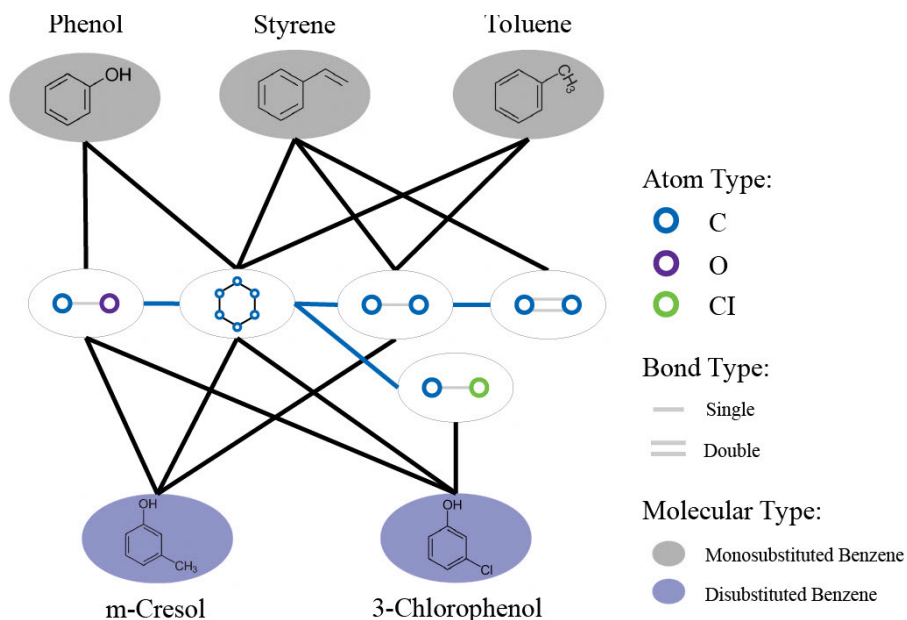
MOTIF VOCABULARY



We build a motif vocabulary by searching all molecular graphs and extract important subgraphs.

- Extract subgraphs from each molecule in the dataset
- Filter the subgraphs and retain recurrent ones
- Remove subgraphs that are not statistically important

HETEROGENEOUS MOTIF GRAPH



Edge weight:

$$A_{ij} = \begin{cases} \text{PMI}_{ij}, & \text{if } i, j \text{ are motifs} \\ \text{TF-IDF}_{ij}, & \text{if } i \text{ or } j \text{ is a motif} \\ 0, & \text{Otherwise} \end{cases}$$

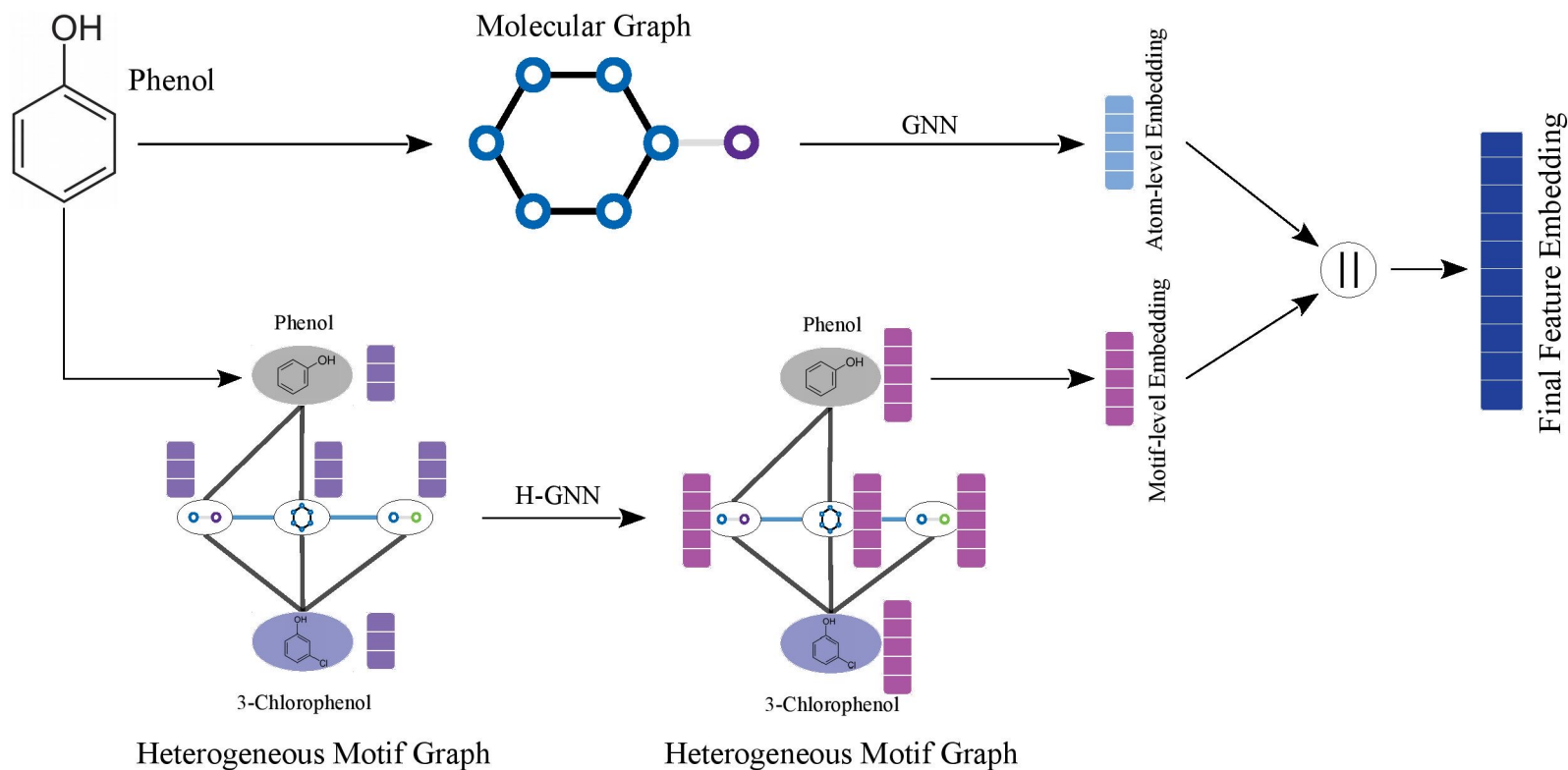
$$\text{TF-IDF}_{ij} = C(i)_j \left(\log \frac{1 + M}{1 + N(i)} + 1 \right)$$

where $C(i)_j$ is the number of times that the motif i appears in the molecule j , M is the number of molecules, and $N(i)$ is the number of molecules containing motif i .

$$\text{PMI}_{ij} = \log \frac{p(i, j)}{p(i)p(j)}$$

Based on the motif vocabulary, we build a heterogeneous graph that contains motif nodes and molecular nodes.

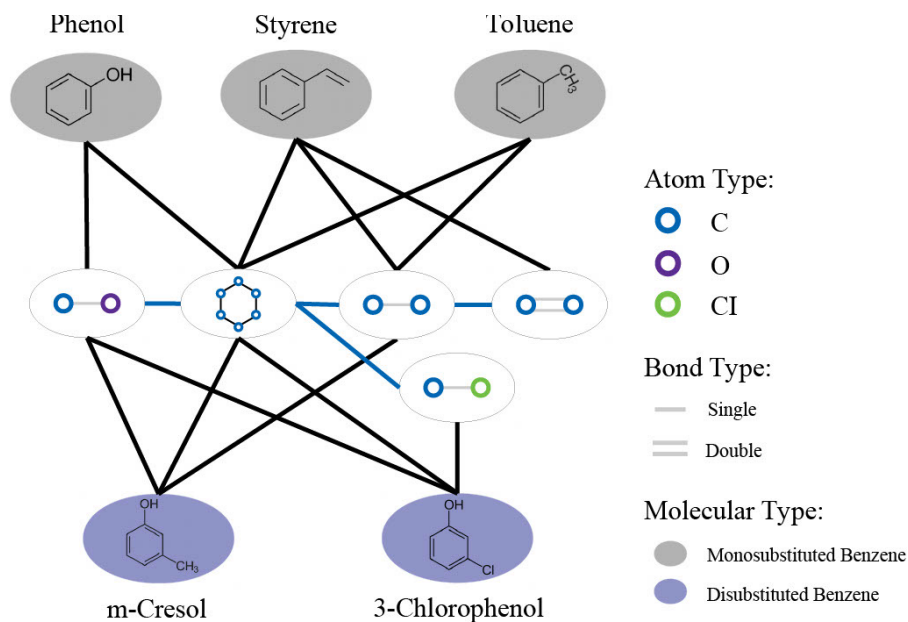
HETEROGENEOUS MOTIF GRAPH NEURAL NETWORKS



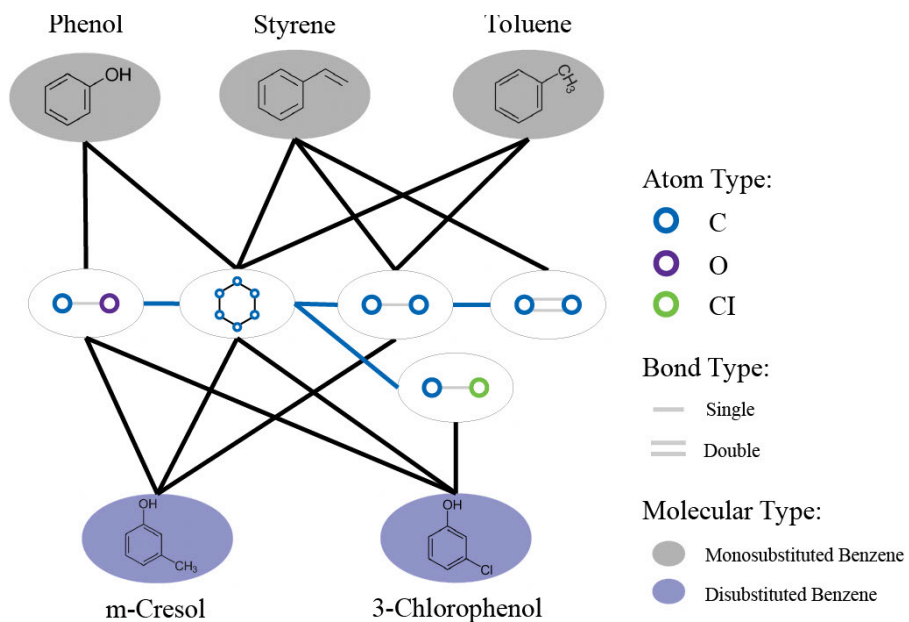
We construct a heterogeneous motif graph to learn both atom-level and motif-level representation simultaneously.

Multi-Task Learning via Heterogeneous Motif Graph

- Construct a heterogeneous motif graph based on shared vocabulary and molecules from all datasets.
- Apply heterogeneous graph neural networks
- Predict for each separate task or dataset



HETEROGENEOUS MOTIF GRAPH

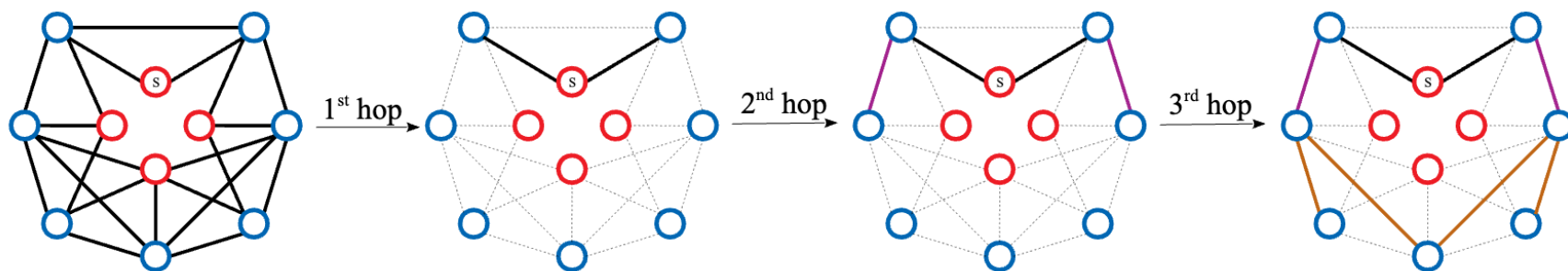


Potential issue:

Efficiency

Based on the motif vocabulary, we build a heterogeneous graph that contains motif nodes and molecular nodes.

Efficient Training via Edge Sampling



In this graph, we have four molecular nodes (red) and seven motif nodes (blue)

- Randomly choose molecular node S as the “starting” node
- In the first hop, keep all edges connecting the node S and motif nodes

EXPERIMENTAL RESULTS

METHODS	PTC	MUTAG	NCI1	PROTEINS	MUTAGENICITY
PatchySAN	60.0 ± 4.8	92.6 ± 4.2	78.6 ± 1.9	75.9 ± 2.8	-
GCN	64.2 ± 4.3	85.6 ± 5.8	80.2 ± 2.0	76.0 ± 3.2	79.8 ± 1.6
GraphSAGE	63.9 ± 7.7	85.1 ± 7.6	77.7 ± 1.5	75.9 ± 3.2	78.8 ± 1.2
DGCNN	58.6 ± 2.5	85.8 ± 1.7	74.4 ± 0.5	75.5 ± 0.9	-
GIN	64.6 ± 7.0	89.4 ± 5.6	82.7 ± 1.7	76.2 ± 2.8	-
PPGN	66.2 ± 6.5	90.6 ± 8.7	83.2 ± 1.1	77.2 ± 4.7	-
CapsGNN	-	86.7 ± 6.9	78.4 ± 1.6	76.3 ± 3.6	-
WEGL	64.6 ± 7.4	88.3 ± 5.1	76.8 ± 1.7	76.1 ± 3.3	-
GraphNorm	64.9 ± 7.5	91.6 ± 6.5	81.4 ± 2.4	77.4 ± 4.9	-
<u>GSN</u>	<u>68.2 ± 7.2</u>	<u>90.6 ± 7.5</u>	<u>83.5 ± 2.3</u>	<u>76.6 ± 5.0</u>	-
OURS	78.8 ± 6.5	96.3 ± 2.6	83.6 ± 1.5	79.9 ± 3.1	83.0 ± 1.1

Motif-info

Ablation Studies on Motif-Motif Interactions

Table 5. Results of a variant without motif-motif interactions and HM-GNN on three datasets.

	PTC	MUTAG	PROTEIN
Variant	76.4±4.4	94.6±4.3	78.3±3.5
HM-GNN	78.8±6.5	96.3±2.6	79.9±3.1

We create a variant of our heterogeneous motif graph by removing all motif-motif interactions from the graph. In this new graph, each molecule use the information of motifs it has.

Ablation Studies on Heterogeneous Motif Graph Neural Networks

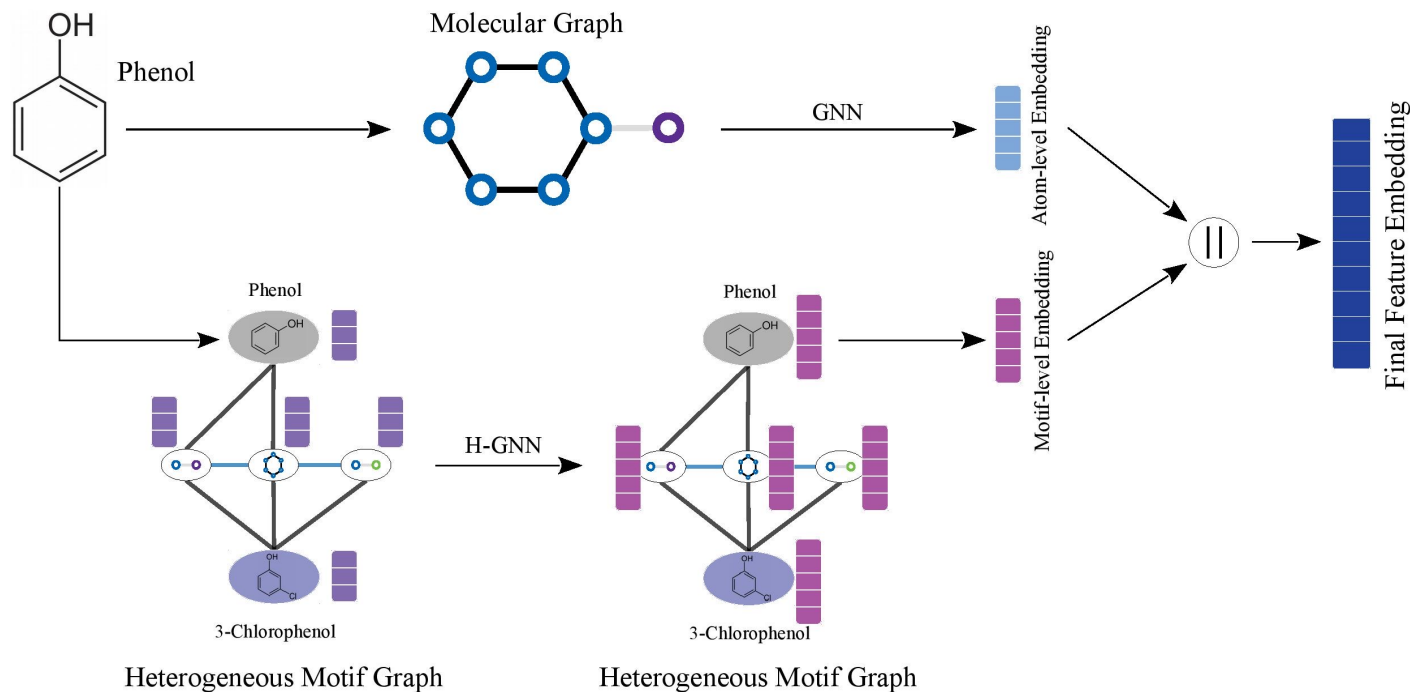


Table 3. Graph classification accuracy (%) of GIN and our model on three datasets: PTC, MUTAG, and PROTEINS.

Models	PTC	MUTAG	PROTEINS
GIN	64.6 ± 7.0	89.4 ± 5.6	76.2 ± 2.8
OURS	78.8 ± 6.5	96.3 ± 2.6	79.9 ± 3.1

Results of Multi-Task Learning on Small Molecular Datasets

Table 4. Results on the PTC dataset with three different training settings. The first row report the performances of only using the PTC dataset. The second row and third row show the results of training on combined vocabularies and datasets with PTC_MM and PTC_FR, respectively. We report the motif vocabulary size (Vocab Size) of the dataset and the Overlap Ratio, which indicates the overlap ratio of motif vocabularies between two datasets. The last three columns represent the performances of using different sizes of training sets. For example, 90% means we use 90% of dataset as the training set and 10% of dataset as the testing set.

Dataset	Vocab Size	Overlap Ratio	90%	50%	10%
PTC	97	-	71.8 \pm 4.1	65.1 \pm 0.8	59.9 \pm 1.9
+ PTC_MM	111	83.5%	76.5 \pm 3.3	69.2 \pm 0.8	66.7 \pm 1.9
+ PTC_FR	110	94.8%	84.3 \pm 3.8	77.3 \pm 0.8	74.0 \pm 1.7

Thank you for listening!

Molecular Representation Learning via
Heterogeneous Motif Graph Neural Networks

Zhaoning Yu and Hongyang Gao