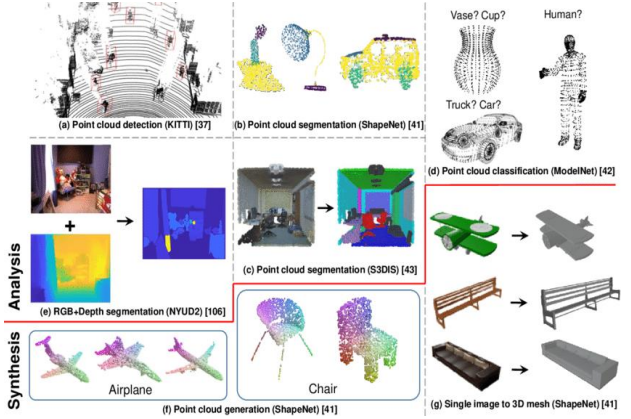
Toward Interpretable Graph Classification via Concept-Focused Structural Correspondence

Tien-Cuong (Alex) Bui
Department of ECE
Seoul National University
Seoul, South Korea

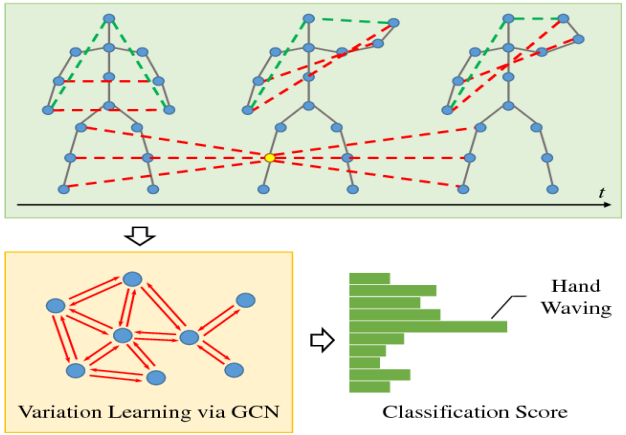
Wen-Syan Li
Graduate School of Data Science
Seoul National University
Seoul, South Korea

Presentation Outline

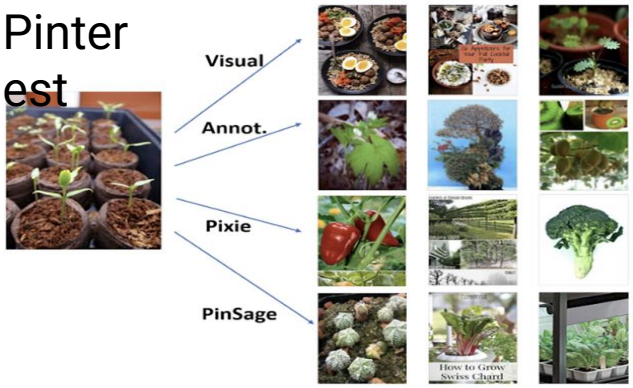
GNNs in Action: Real-world Use Cases



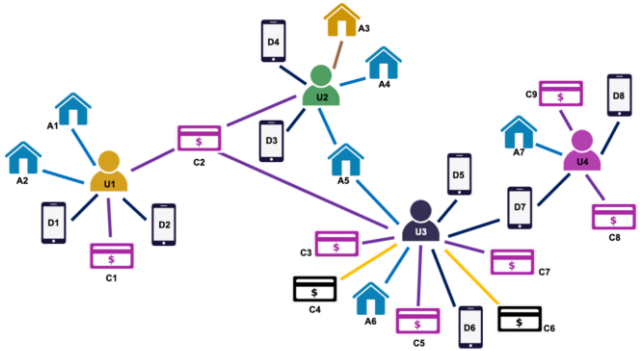
3D Reconstruction



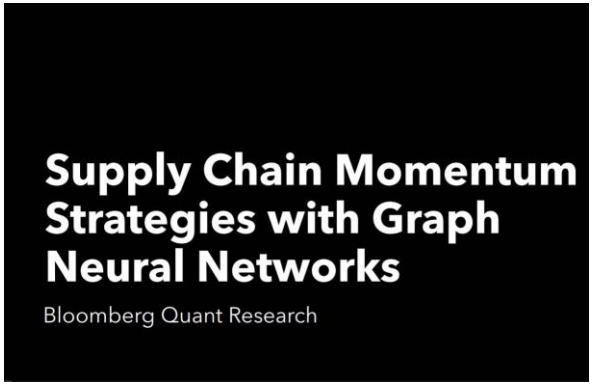
Skeleton-based Action Recognition



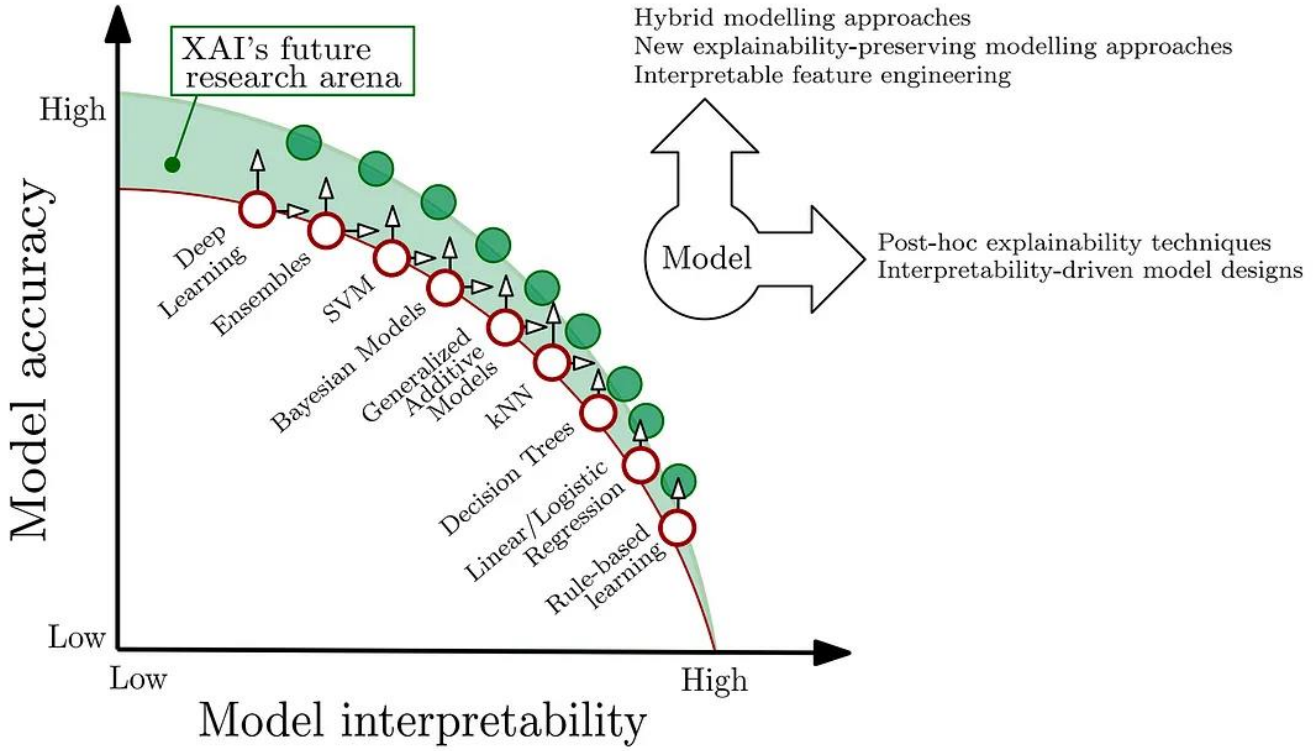
Recommender Systems



Fraud Detection



Interpretability vs Accuracy Trade-off



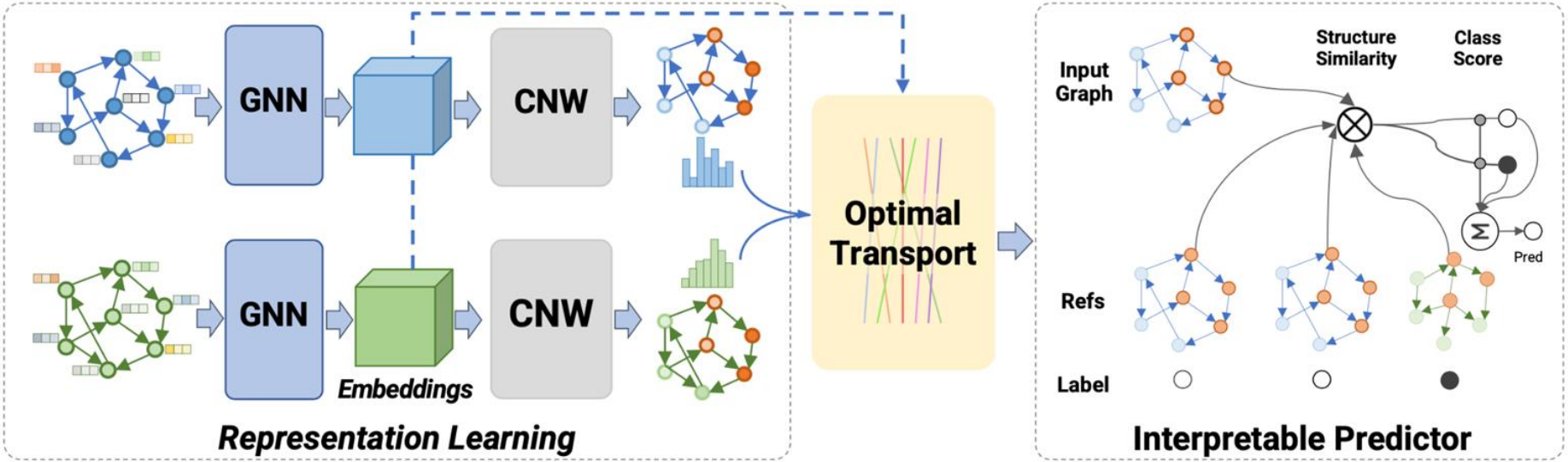
https://medium.com/data-folks-indonesia/overview-interpretable-machine-learning-cb0b4f2e01dd

Shortcomings of XAI Methods for Graphs

- **Feature attribution**
  - Traditional methods struggle with the relational nature of graphs
  - Focus on individual nodes/edges might miss complex interactions
- **Example-based approach**
  - Identifying relevant examples in large graphs is challenging
  - May not reveal the underlying logic behind a model's decision
- **Post-hoc GNN Explanations**
  - Require additional computation costs
  - Relying solely on the model's outputs to generate explanations
  - Inaccurate explanations, particularly for complex models
- **Self-explainable GNNs**
  - Ignore existing GNN architectures
  - Model-specific and difficult to generalize

Framework Overview

- Inspired by **natural interpretability of KNN**
- **Phase 1:** Learning concept-focused graph representations
- **Phase 2:** Infer graph classes based on graph structure similarity
  - Measuring graph similarity via **concept-focused optimal transport** distance

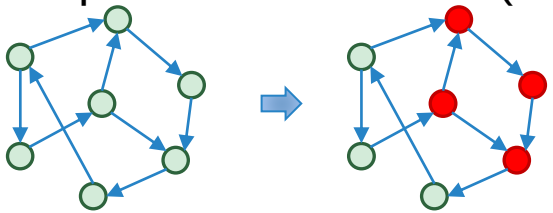


Learning concept-based graph representations

Inference phase

Phase 1: Concept-focused Graph Representation Learning

- Goal: extracting frequent substructures (concepts) signifying specific outcomes



- Formulate the optimization problem based on graph information bottleneck

$$\max_{\tilde{\mathcal{G}} \subset \mathcal{G}} I(\hat{Y}, \tilde{\mathcal{G}}) - \alpha I(\mathcal{G}, \tilde{\mathcal{G}}) + \beta I(\hat{Y}, \mathcal{G})$$

Subgraph extraction constraint

- *I ~ mutual information between 2 variables*
  - Optimizing I(Y, G) via cross-entropy loss
  - Optimizing I(G, G~) via f-divergence of KL-divergence (Donsker and Varadhan)

Yu et al., “Graph information bottleneck for subgraph recognition”

Phase 2: Interpretable Prediction

- **Step 1:** Collect concept-based graph representations
  - Execute representation module over all training graphs
  - Manage representations in a vector database
- **Step 2:** Two-stage reference selection
  - Euclidean-based retrieval, followed by reranking with structure similarity metric

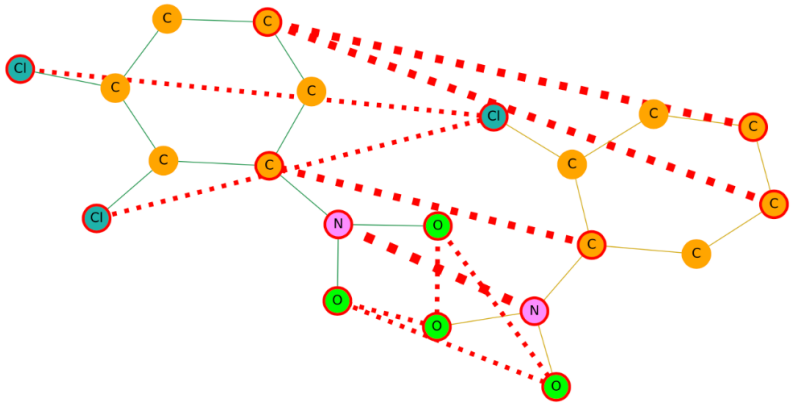
$$\mathcal{R}_{\mathcal{G}} = \text{Structure\_Rank}(\mathcal{R}^e_{\mathcal{G}}, K) \quad \text{s.t.} \quad \mathcal{R}^e_{\mathcal{G}} = \text{KNN}(h_{\mathcal{G}_s}, \alpha K)$$

- **Step 3:** Infer prediction based on structure similarity to reference concepts

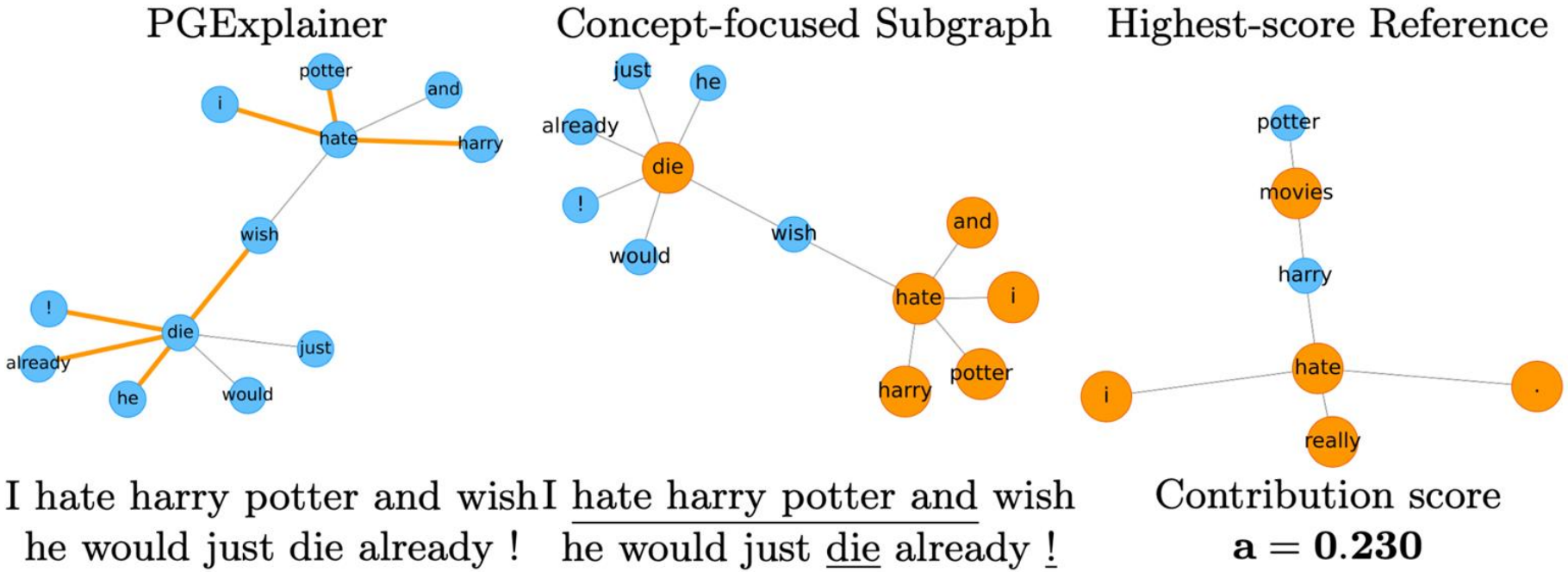
$$P(\hat{Y}|\mathcal{G}, \mathcal{R}_{\mathcal{G}}) = \sum_{i=1}^{K} a(\mathcal{G}, R_i) Y_i \quad \text{s.t} \quad a(\mathcal{G}, R_i) = \text{softmax}(s_{\text{sc}}(\mathcal{G}, R_i))$$

Graph Structure Similarity

- **Goal:** Measuring structure similarity between two graphs
- **Assumption:** node embeddings hold structure information
- **Relaxation:** Graph structure similarity ~ distance between two distributions
- **Problem:** Finding an optimal transport flow matrix ***T**** between two distributions
- **Solution:**
  - $D = (d_{ij}) \in R^{N \times N}$ ~ a ground distance matrix
    - d_{ij} ~ Euclidean distance between two interconnected nodes
  - If d_{ij} is large, t_{ij} must be small, otherwise
  - Solve the optimization problem via Sinkhorn algorithm



Explanation Construction



An example of a prediction's explanation

DATASETS

- Conducted experiments with 5 graph classification datasets
- Using 10-fold cross-validation with 8:1:1 splitting strategy

Dataset Name	#Graphs	#Avg vertices	#Avg edges	#Features	#Labels
Mutag	188	17.93	19.79	7	2
Proteins	1113	39.06	72.82	29	2
IMDB-Binary	1000	19.77	96.53	271	2
DD	1178	284.32	715.66	89	2
Twitter	6940	11.9K	20.10	768	3

K: 10^3

BASELINES

- Compared with 2 groups of baselines & 4 GNN backbones
- GNN backbones:
  - **GCN:** a fundamental GNN model based on spectral graph concept
  - **GraphSage:** an inductive learning framework leveraging graph convolutions and neighbor sampling to generate node embeddings
  - **GIN:** employing learnable aggregation functions to increase expressive power
  - **GAT**: modelling interactions between nodes
- 2 groups of baselines:
  - **Group 1:** training backbones with cross-entropy loss function
  - **Group 2:** training backbones with graph information bottleneck theory

Yu et al., “Graph information bottleneck for subgraph recognition”

Accuracy Comparison

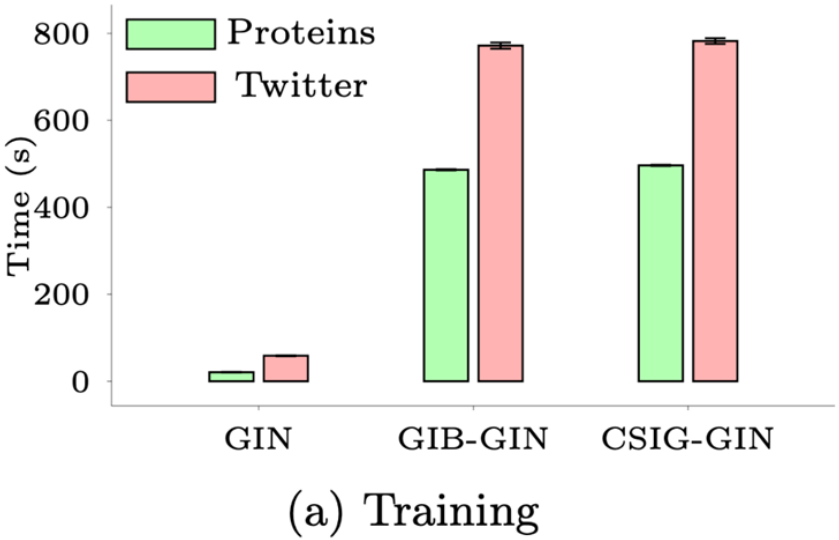
- Our proposed method boosts predictive performance of GNNs
- Outperformance two groups of baselines

	Backbone	Mutag	Proteins	IMDB	DD	Twitter
Backbone Training	GCN	0.718	0.714	0.710	0.715	0.642
	GraphSage	0.730	0.694	0.715	0.743	0.636
	GIN	0.862	0.750	0.726	0.699	0.651
	GAT	0.750	0.672	0.726	0.699	0.664
GIB Training	GCN	0.772	0.731	0.726	0.765	0.513
	GraphSage	0.750	0.699	0.720	0.772	0.546
	GIN	0.841	0.721	0.702	0.729	0.630
	GAT	0.771	0.684	0.717	0.698	0.505
Interpretable Predictors	GCN	0.846	0.706	0.702	**0.785**	0.678
	GraphSage	0.862	0.739	0.728	0.778	0.658
	GIN	**0.877**	**0.757**	0.724	0.750	0.633
	GAT	0.798	0.738	**0.729**	0.769	**0.683**

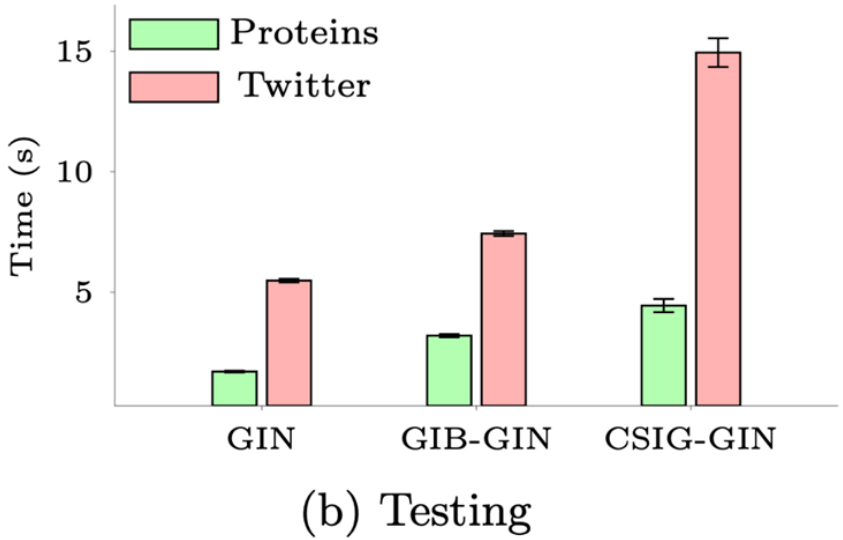
GIB: Graph information bottleneck

Execution Time Evaluation

- Representation module focus on discovering patterns and concepts in training
- Calculating structure similarity is the main bottleneck in inference



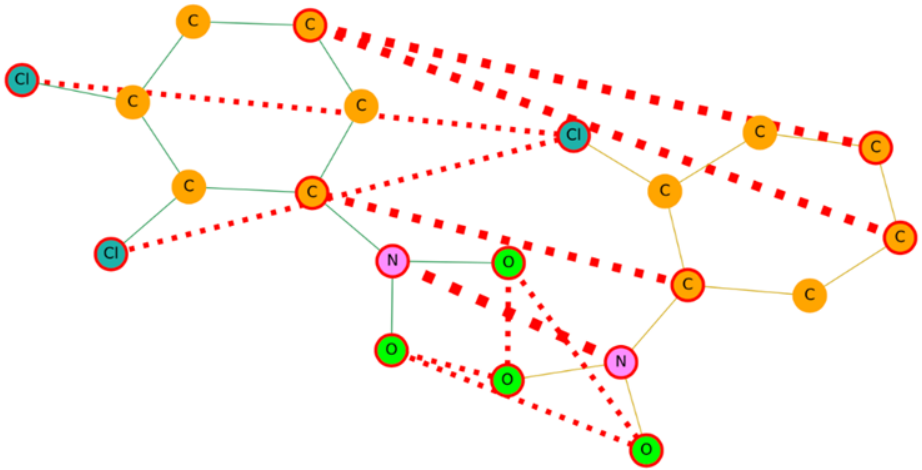
(a) Training



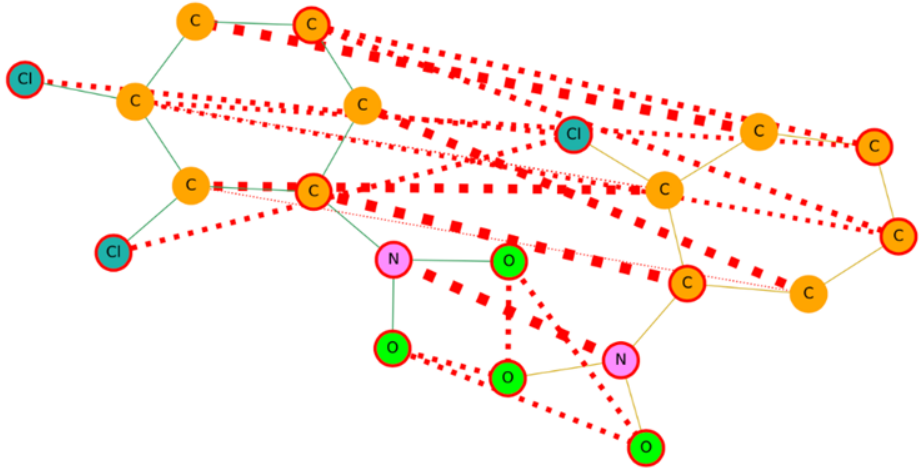
(b) Testing

Qualitative Evaluation

- Concept-focused node initialization yields clearer visualization
- Only focus on most similar nodes between two graphs



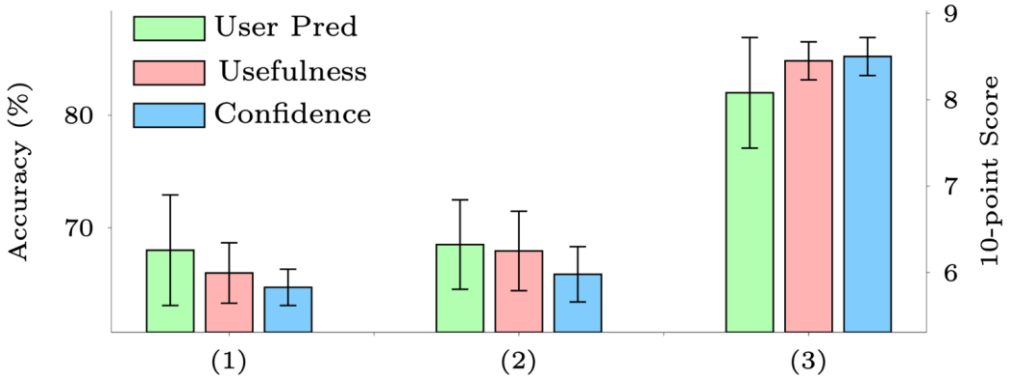
(a) Concept-based Correspondence



(b) Uniform-based Correspondence

User Evaluation of Prediction Explanations

- **Goal:** Assessing user perception of different explanation modalities
- **How:** Organizing a user study with 20 participants
  - Predicting model outcomes given different explanation types
- **Notable Insights:**
  - Presenting only subgraphs had a limited impact on user comprehension and confidence
  - Concept-based references improved user understanding and consensus with the model
  - Incorporating diverse information types enhances user comprehension



(1) PG-Explainer subgraph
(2) Concept-focused subgraph
(3) Concept-focused subgraph and references with similarity scores

Conclusion & Future Work

- Introduced a concept-based approach to interpretable graph classification
- Structural similarity with Earth Mover's Distance (EMD) enhances interpretability
- Non-parametric prediction model yields clear explanations
- Efficient computation with a dual-phase distance strategy
- Explanations tailored for diverse user needs
- Comprehensive evaluation and user study confirm effectiveness
- **Future Work**:
  - Incorporating human-knowledge constraints into concept discovery
  - Organizing the concept corpus hierarchically for efficient exploration
  - Introducing interactive features with a user-friendly interface to enhance interpretability and usability

THANK YOU !

QUESTIONS