Size-Invariant Graph Representations for Graph Classification Extrapolations

Beatrice Bevilacqua*, Yangze Zhou*, Bruno Ribeiro

Purdue University

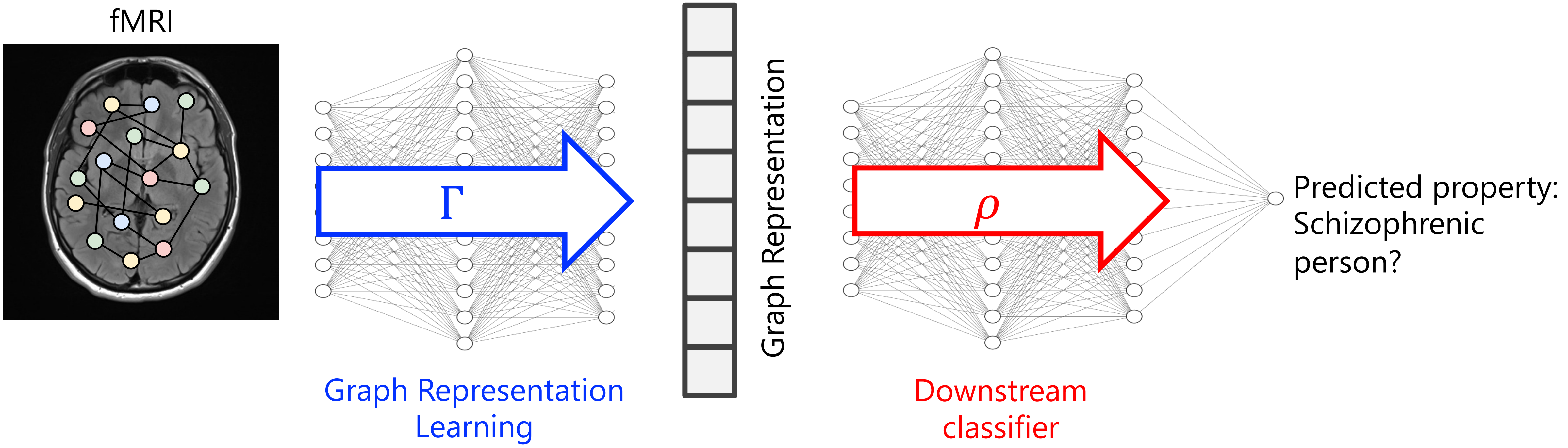
ICML 2021

*Equal contribution

This work focuses on out-of-distribution (OOD) extrapolations in Graph Representation Learning

Toolbox:
- Causality
- Graph limits
- Graph Neural Networks

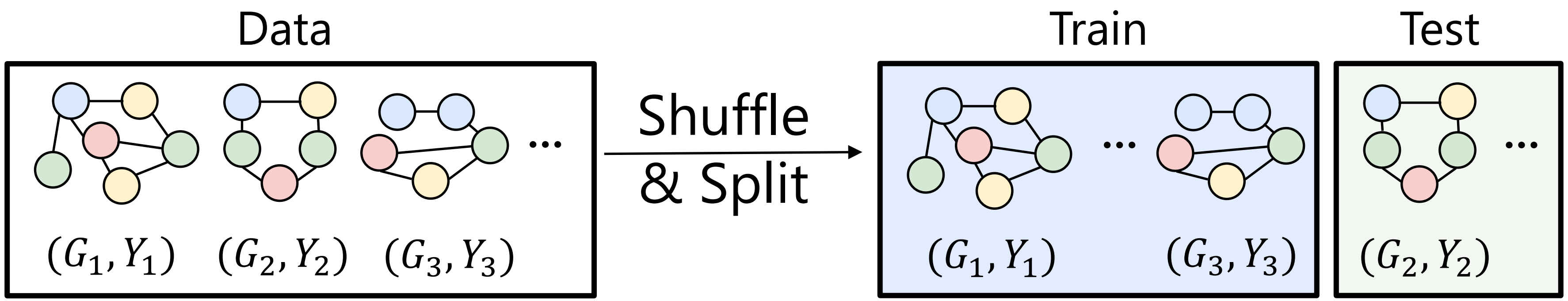
Graph Classification Tasks



Current Graph Classification Approach

Graph Representation Learning generally assumes:

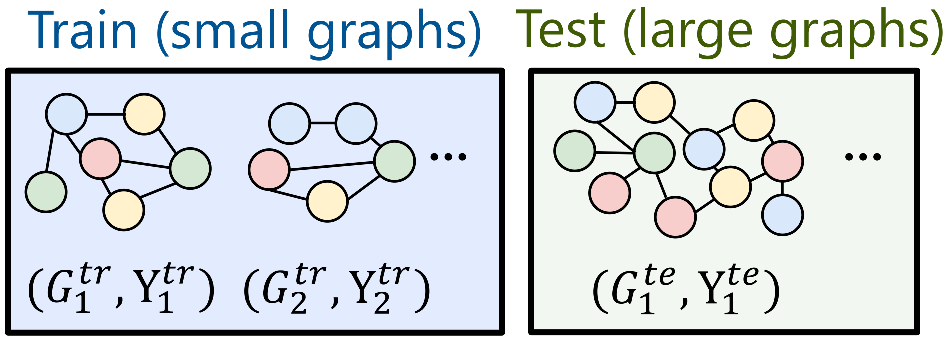
Train distribution = Test distribution



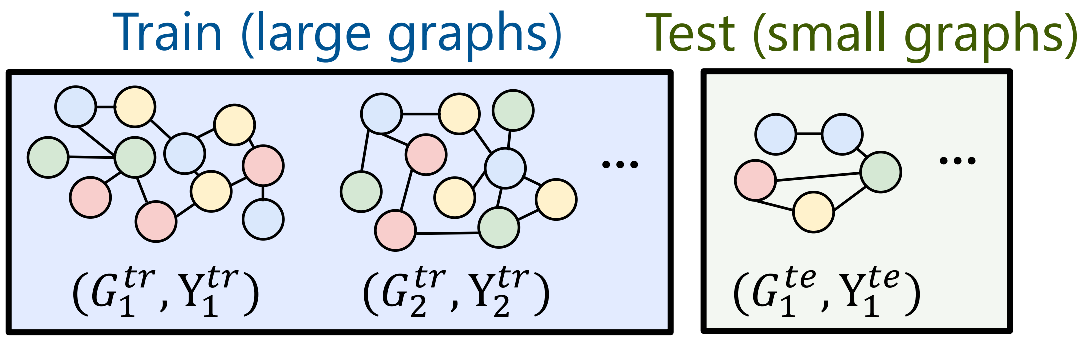
What if test data were out of distribution (OOD)?

Extrapolation to Different Graph Sizes

What if train has **small** graphs but test has **large** graphs?



What if train has **large** graphs but test has **small** graphs?



Bevilacqua, Zhou, Ribeiro, ICML 2021, Size-Invariant Graph Representations for Graph Classification Extrapolations

Size Extrapolation with GNNs?

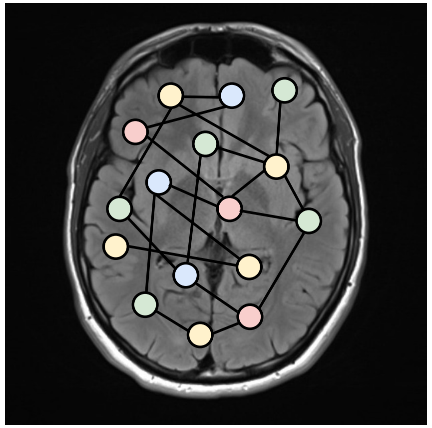
Purdue University

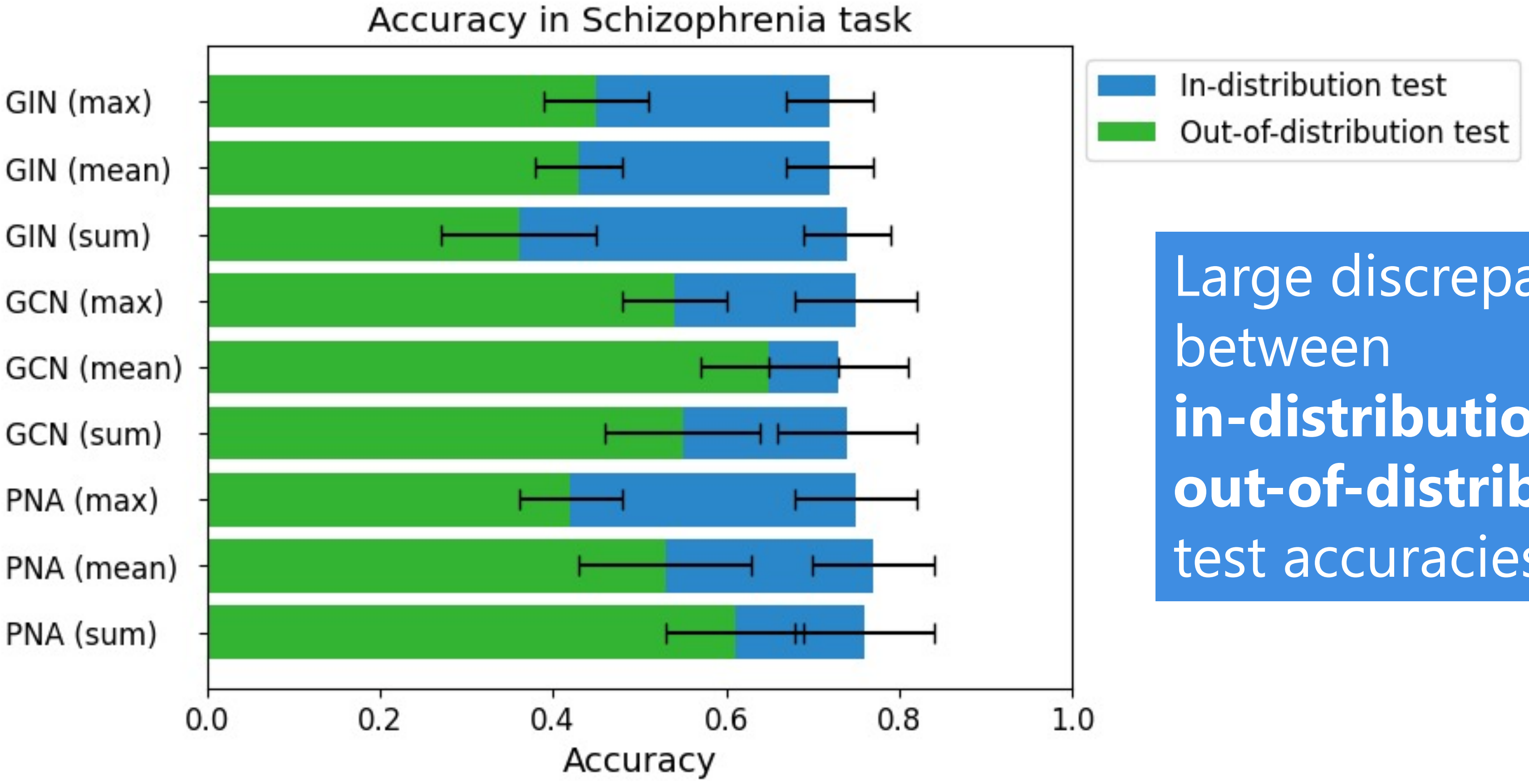
Do Graph Neural Networks (GNNs) extrapolate?

⇒ GNNs can be applied to graphs of any size

⇒ But may not extrapolate between **small (train)** and **large (test)** graphs:

Schizophrenia task





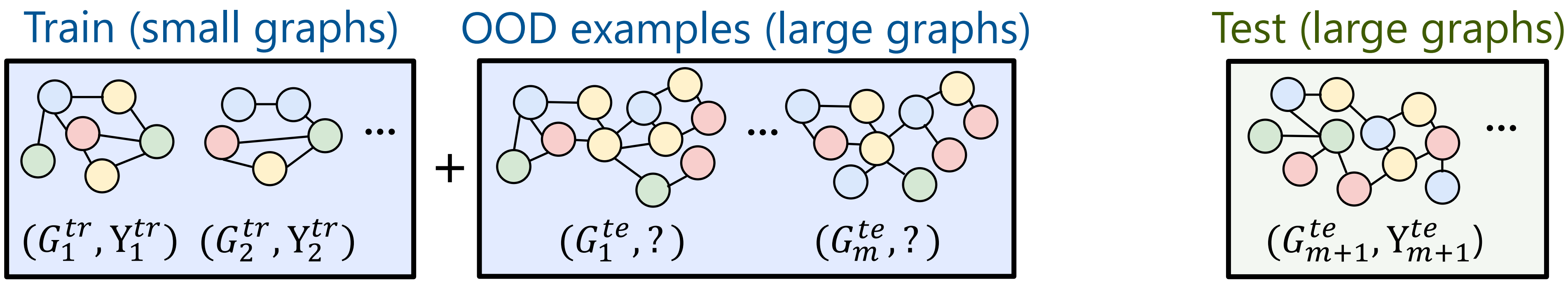
Large discrepancy between **in-distribution** and **out-of-distribution** test accuracies

How to Extrapolate in Graph Classification Tasks?

How do we extrapolate beyond the training distribution?

If OOD examples available, **data-driven methods** work:

- Domain Adaptation
- Covariate Shift Adaptation
- Few-shot Learning
- Data Augmentation
- Invariant Risk-Minimization (IRM)*



How to Extrapolate in Graph Classification Tasks?

Data-driven methods:

Pros

- Can use existing GNN methods
- Don't assume a mechanism for distribution shift

Cons

- Must have OOD examples during training

What if no access to OOD data?

- Must define a causal mechanism

Next: Observational vs Causal (Counterfactual) modeling

Purdue University

Why are Causal Mechanisms Needed for Extrapolations without OOD Data?

Graph Representation Learning is Observational

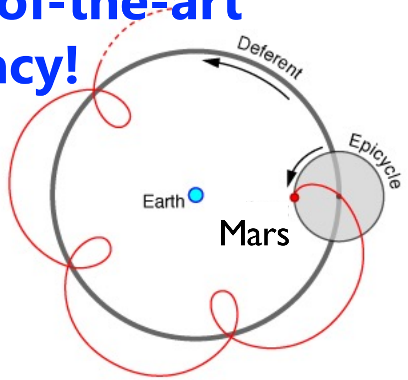
PURDUE UNIVERSITY

Historical analogy to Graph Representation Learning methods:

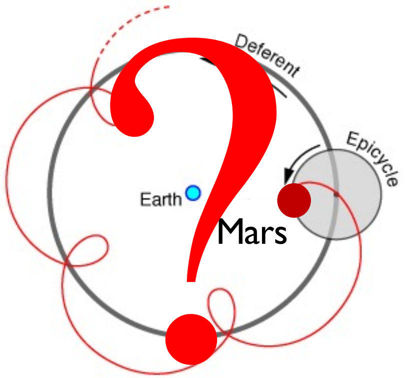
- Ptolemaic geocentric model of planetary motion
  - Very **accurate** to predict positions **observationally**
  - **Cannot** predict positions in new **scenarios**

New scenario:
What would happen if
Mars became 10x more massive?

State-of-the-art accuracy!

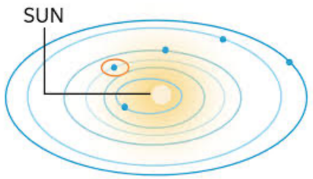


Mars gains mass →



img credits: wikipedia

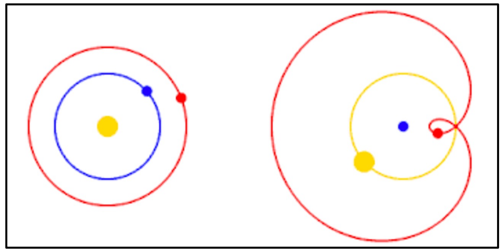
Interpretable model
cannot predict new scenarios



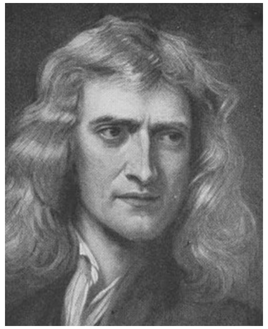
Lesson:
Occam's razor & interpretability
≠ out-of-distribution extrapolation

Extrapolations to New Scenarios

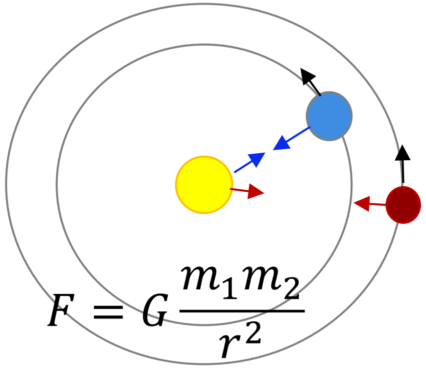
- Observational predictions can be purely data-driven



- Predicting new scenarios (larger and smaller mass) without OOD examples requires a **mechanism**



New scenario:
What would happen if
Mars became 10x more massive?

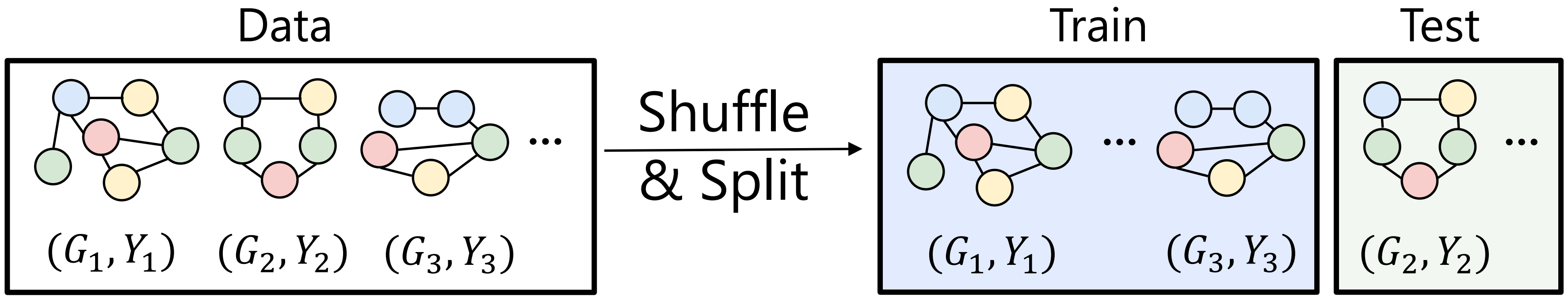


Size Extrapolations on Graphs

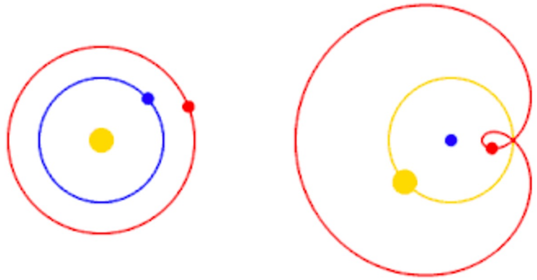
Differences between Observational and Counterfactual Tasks

Observational Task:

Predicting unseen examples of training distribution

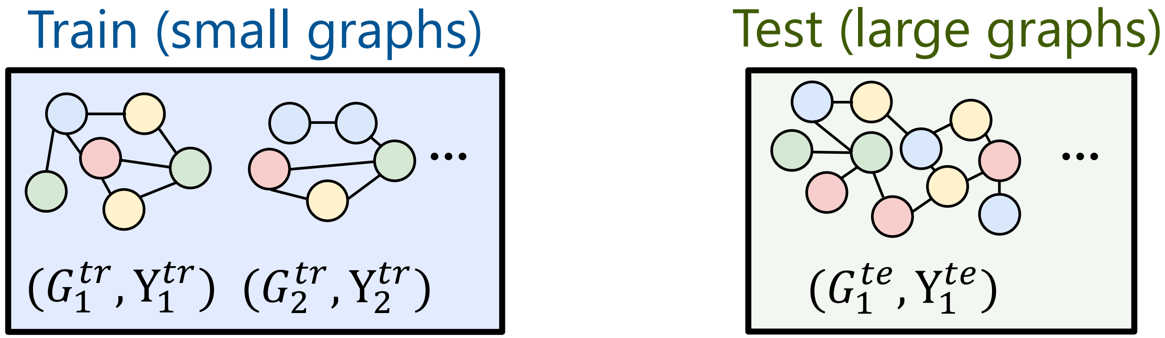


Planetary Motion Equivalent

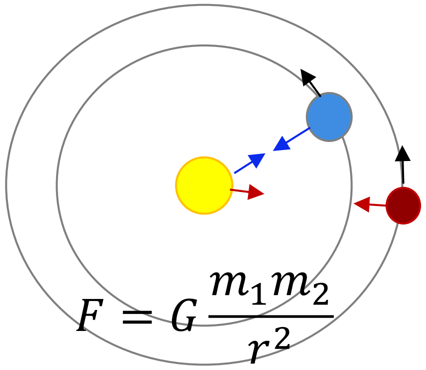


Counterfactual Task (since we have no access to test data)**:**

What would be the label of a graph if it were larger?



or vice-versa



Reminder of talk:

What would be the labels if the graphs were larger?

To Infinity and Beyond…

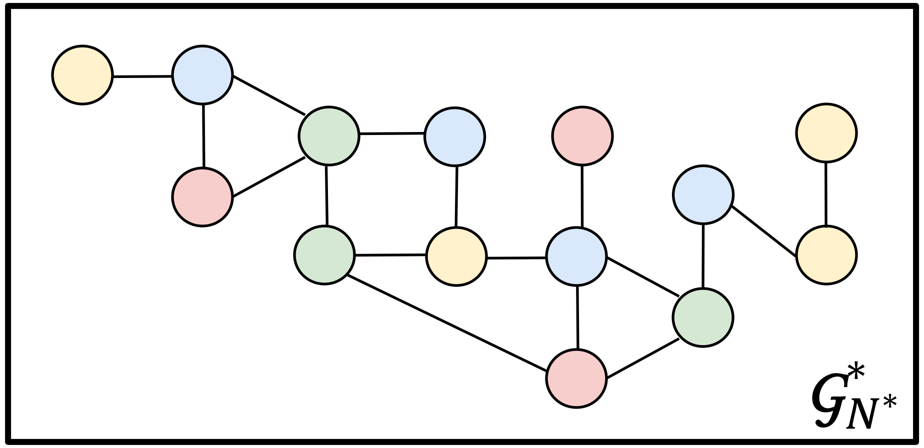
Q: What would be the label if the graph were infinitely large?

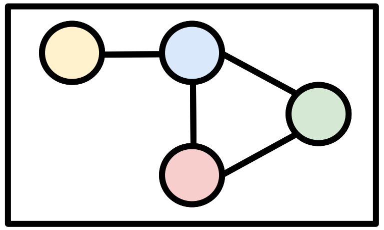
$$N \to \infty$$

Lovász Graph Limits

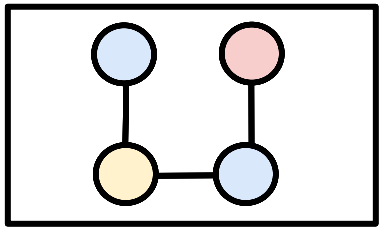
- What graph property is invariant as graphs become larger?
  - Lovász & Szegedy (2006) shows:
    - Density of induced subgraphs of a dense random graph converges as $N \to \infty$

$\mathcal{G}^*_{N^*}$ can be train $\mathcal{G}^{tr}_{N^{tr}}$ or test $\mathcal{G}^{te}_{N^{te}}$ graph





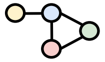
Count = 2

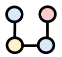


Count = 1

Induced k-sized subgraph density

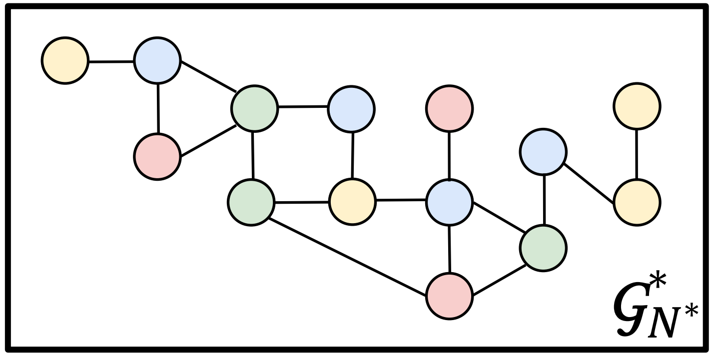
$$\mathrm{t}_{\mathrm{ind}}(F_k, \mathcal{G}^*_{N^*}) = \frac{\mathrm{ind}(F_k, \mathcal{G}^*_{N^*})}{N^*! \,/\, (N^* - k)!}$$

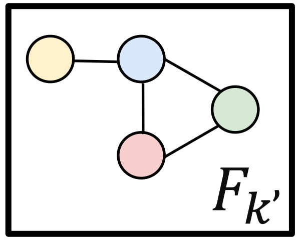
$\mathrm{t}_{\mathrm{ind}}($  $, \mathcal{G}^*_{N^*}) = \dfrac{2}{14! \,/\, (14 - 4)!}$

$\mathrm{t}_{\mathrm{ind}}($  $, \mathcal{G}^*_{N^*}) = \dfrac{1}{14! \,/\, (14 - 4)!}$

What if we constructed a graph representation from subgraph densities?

Graph Representation based on Densities





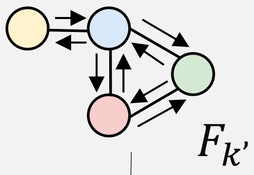
New graph representation

Induced subgraph density

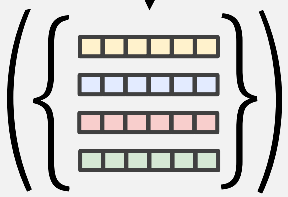
$$\Gamma_{\text{GNN}}(\mathcal{G}^*_{N^*}) = \sum_{F_{k'} \in \mathcal{F}_{\leq k}} \text{t}_{\text{ind}}(F_{k'}, \mathcal{G}^*_{N^*})\text{READOUT}_{\Gamma}(\text{GNN}(F_{k'}))$$

GNN-representation of $F_{k'}$

$\text{GNN}(F_{k'})$:

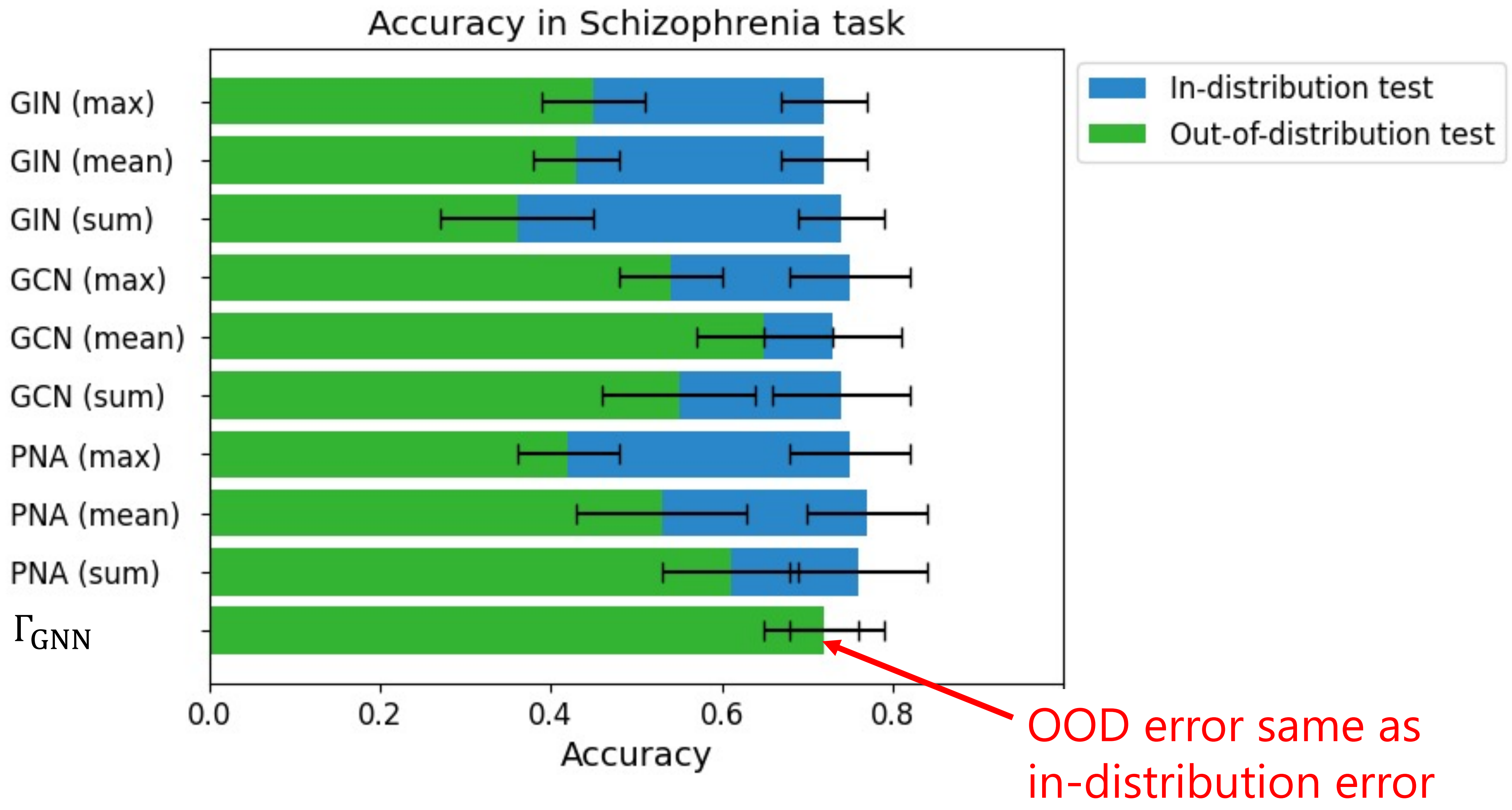


READOUT_{Γ}



OOD Error in Schizophrenia Task

- Can subgraph density representation Γ_{GNN} extrapolate OOD?



Bevilacqua, Zhou, Ribeiro, ICML 2021, Size-Invariant Graph Representations for Graph Classification Extrapolations

Theory

Understand why Γ_{GNN} can OOD extrapolate

A Most Expressive Representation

Preliminary: **1-hot encoded graph representations are most-expressive**

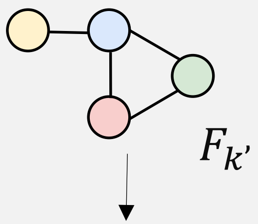
First: replace the GNN-representation of $F_{k'}$ with a 1-hot encoded representation

Graph representation

Induced subgraph density

$$\Gamma_{1-\text{hot}}(\mathcal{G}^*_{N^*}) = \sum_{F_{k'} \in \mathcal{F}_{\leq k}} \mathrm{t}_{\text{ind}}(F_{k'}, \mathcal{G}^*_{N^*}) \mathbb{1}_{\text{one-hot}}\{F_{k'}, \mathcal{F}_{\leq k}\}$$

One-hot encoded identifier of $F_{k'}$



(0,0, .., 1, 0, ..)

Size-Invariant Representation

Theorem 1 (informal): Approximately size-invariant representations
Under certain conditions (explained later), the change in graph representation between train and counterfactual test graph is upper bounded by k and graph sizes (in train and test):

$$\mathrm{P}\left(\|\Gamma_{1-\mathrm{hot}}\left(\mathcal{G}^{tr}_{N^{tr}}\right) - \Gamma_{1-\mathrm{hot}}\left(\mathcal{G}^{te}_{N^{te}}\right)\|_\infty > \epsilon\right) \leq 2|\mathcal{F}_{\leq k}|\left(\exp\left(-\frac{\epsilon^2 N^{tr}}{8k^2}\right) + \exp\left(-\frac{\epsilon^2 N^{te}}{8k^2}\right)\right)$$

Training graph — Counterfactual test graph

- Proof relies on Lovász graph limits (formal definition in paper)

Note that Γ_{GNN} is less expressive (more invariant) than $\Gamma_{1-\mathrm{hot}}$

Effects of Invariant Representations

PURDUE UNIVERSITY

- Why are we interested in invariant representations?

Proposition 1 (informal): Effect of invariant representations
Consider:
- Γ : A permutation invariant graph representation
- ρ : A downstream classifier

In-distribution generalization error : $\forall y \in Y$, for some $\epsilon, \delta \geq 0$

$$\mathrm{P}(|\mathrm{P}(Y = y \mid \mathcal{G}^{tr}_{N^{tr}}) - \rho\big(y, \Gamma(\mathcal{G}^{tr}_{N^{tr}})\big)| \leq \epsilon) \geq 1 - \delta$$

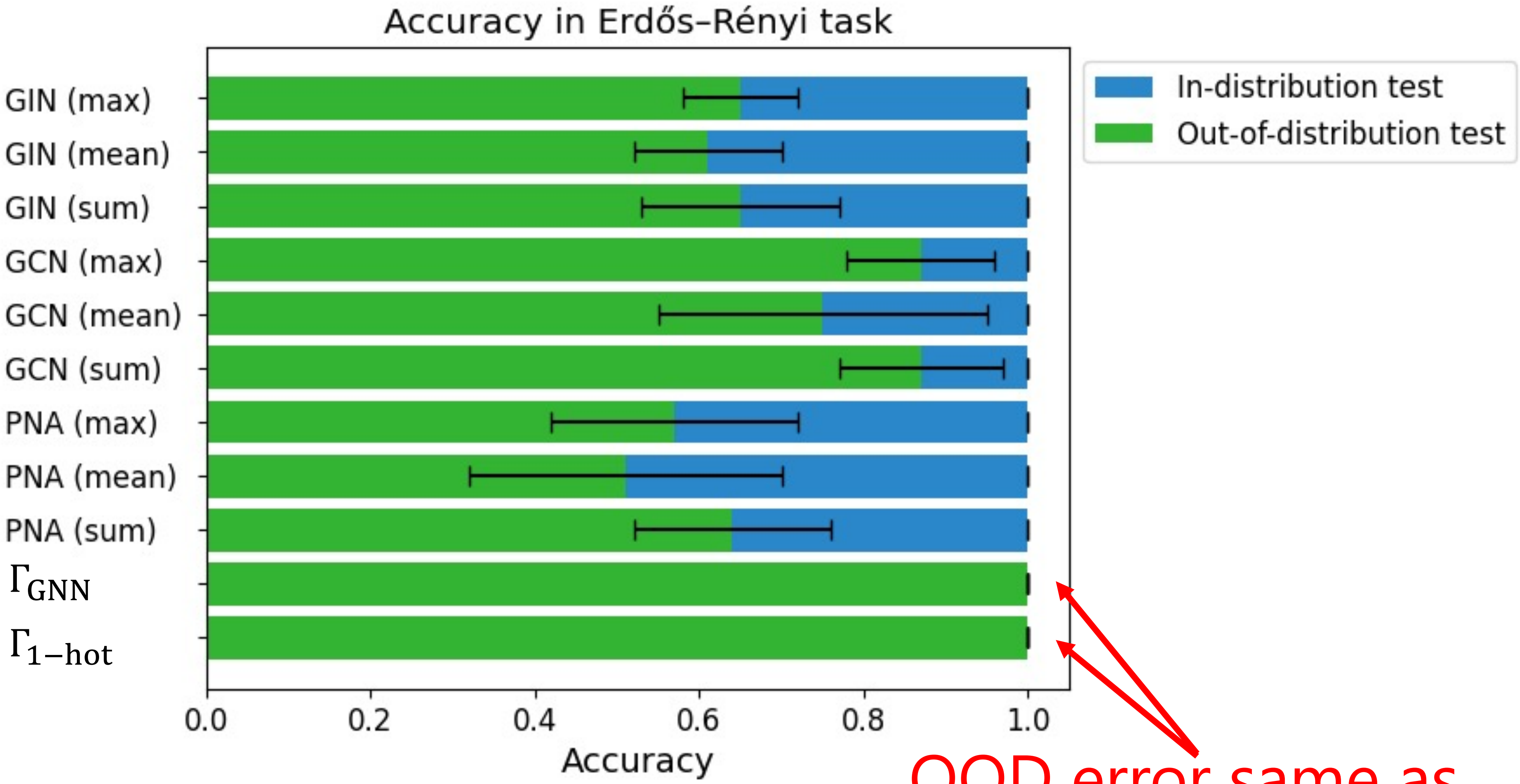
If Γ is OOD-invariant then test error is the same

$$\mathrm{P}(|\mathrm{P}(Y = y \mid \mathcal{G}^{te}_{N^{te}}) - \rho\big(y, \Gamma(\mathcal{G}^{te}_{N^{te}})\big)| \leq \epsilon) \geq 1 - \delta$$

A size-invariant representation has same error
in-distribution and **out-of-distribution**

Bevilacqua, Zhou, Ribeiro, ICML 2021, Size-Invariant Graph Representations for Graph Classification Extrapolations

Erdős–Rényi Task Example



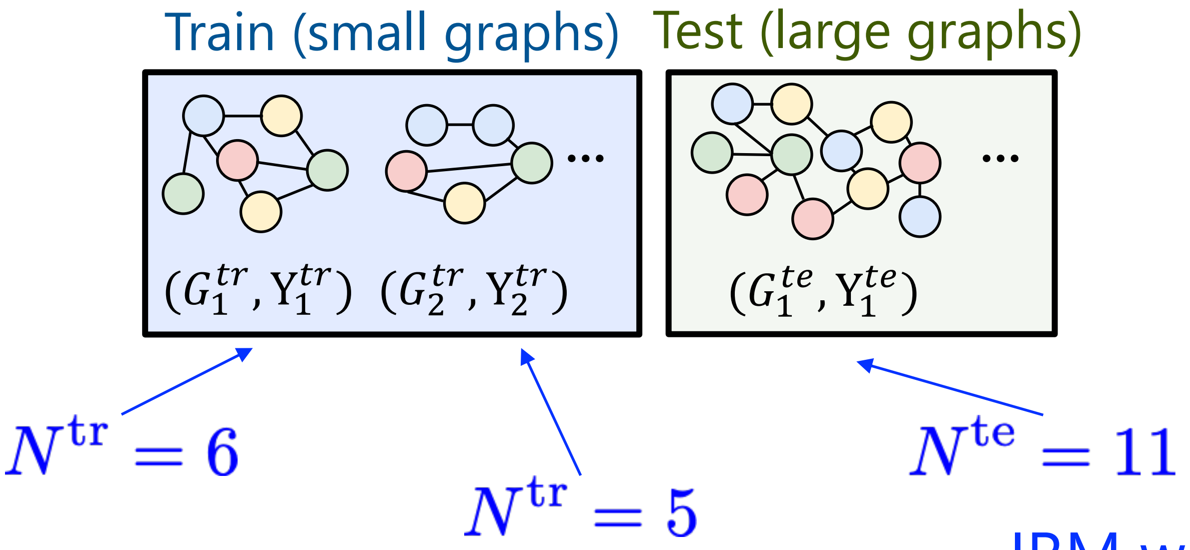
OOD error same as in-distribution error

Invariant Risk Minimization (IRM)

PURDUE UNIVERSITY

IRM (Arjovskj et al., 2019) aims to learn an invariant representation.
However:

- 🚫 no guarantees if representation is nonlinear (e.g., GNN)
- 🚫 not applicable if training graphs have same size
- 🚫 not invariant if OOD support ≠ training support



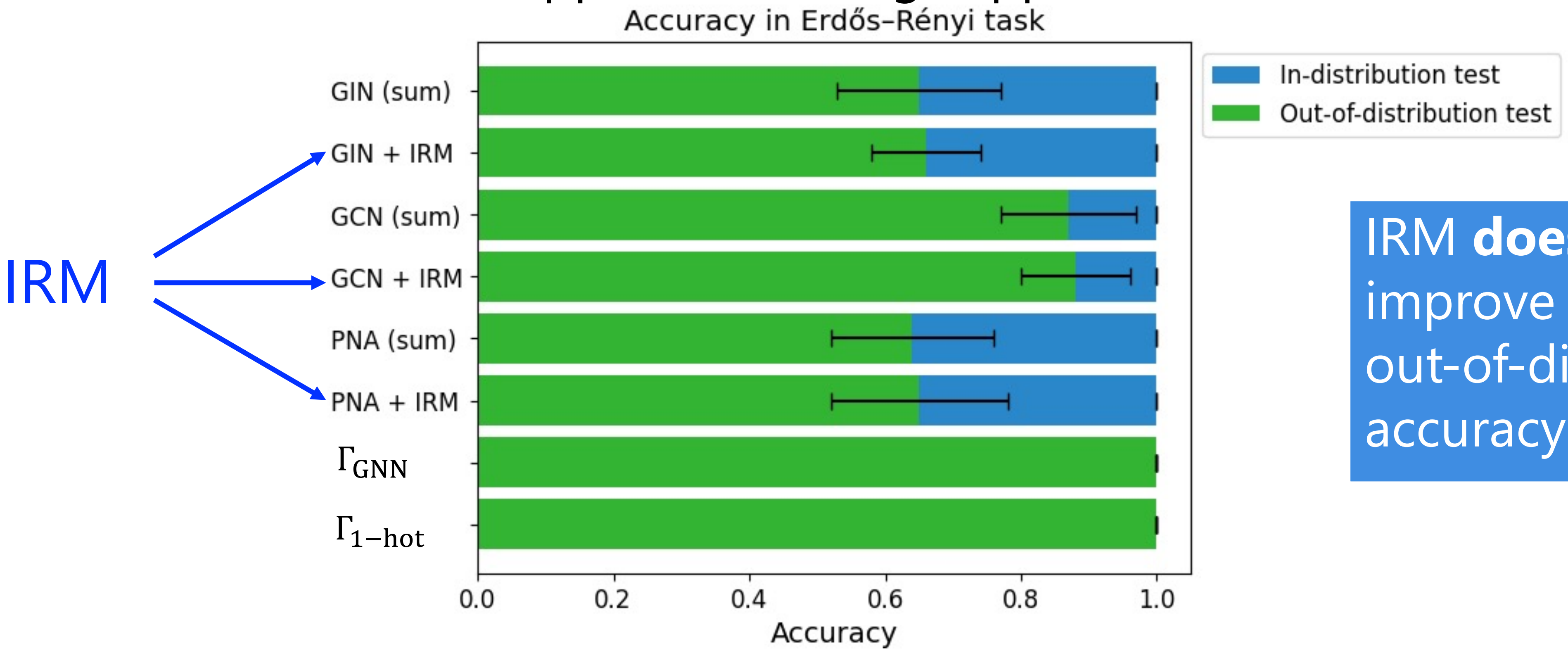
IRM will not work if 11 is never seen in train

Invariant Risk Minimization (IRM)

IRM (Arjovskj et al., 2019) aims to learn an invariant representation.

However:

- 🚫 no guarantees if representation is nonlinear (e.g., GNN)
- 🚫 not applicable if training graphs have same size
- 🚫 not invariant if OOD support ≠ training support

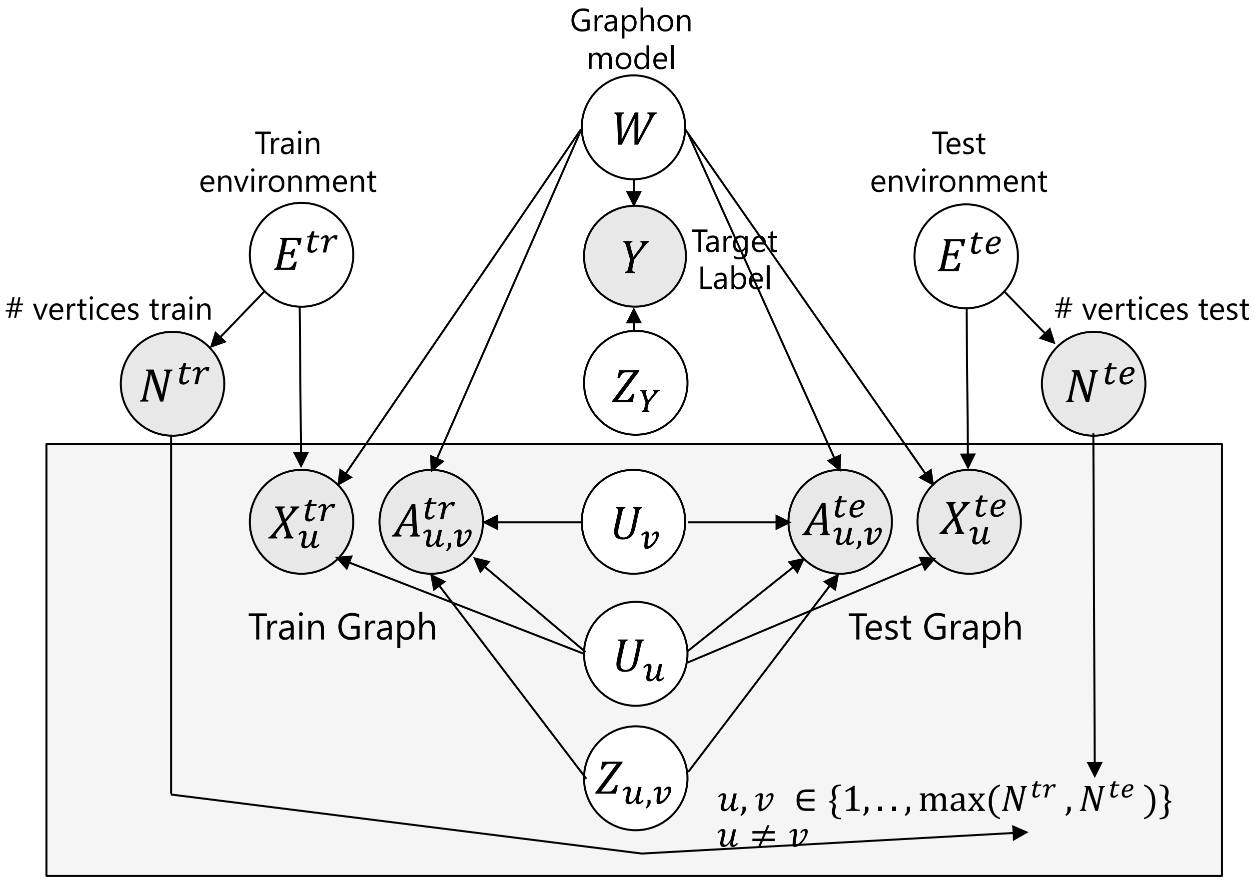


IRM **does not** improve out-of-distribution accuracy

Causal Mechanism Assumed by Theorem 1 & Proposition 1

Causal Mechanism Assumed by **Theorem 1 & Proposition 1**

- *Structural Causal Model:*
  - Graph label Y is a function of the graph model W + some random noise
  - Graph size N^{tr} (N^{te}) is a function of "environment" E^{tr} (E^{te}) only
  - Train (test) graphs are generated by W and E^{tr} (E^{te}) with same random noises

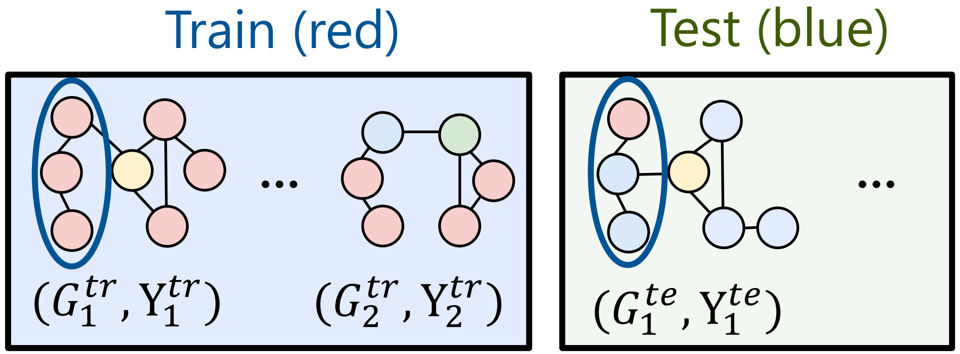


Improving OOD extrapolation of vertex attributes

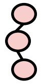
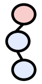
Symmetry Regularization

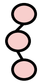
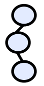
PURDUE UNIVERSITY

- What if OOD shift in attribute distribution?



- Attribute symmetry regularization for representation Γ_{GNN}:

Loss $+ \lambda \| \text{READOUT}_\Gamma(\text{GNN}($  $)) - \text{READOUT}_\Gamma(\text{GNN}($  $)) \|$

$+ \lambda \| \text{READOUT}_\Gamma(\text{GNN}($  $)) - \text{READOUT}_\Gamma(\text{GNN}($  $)) \|$

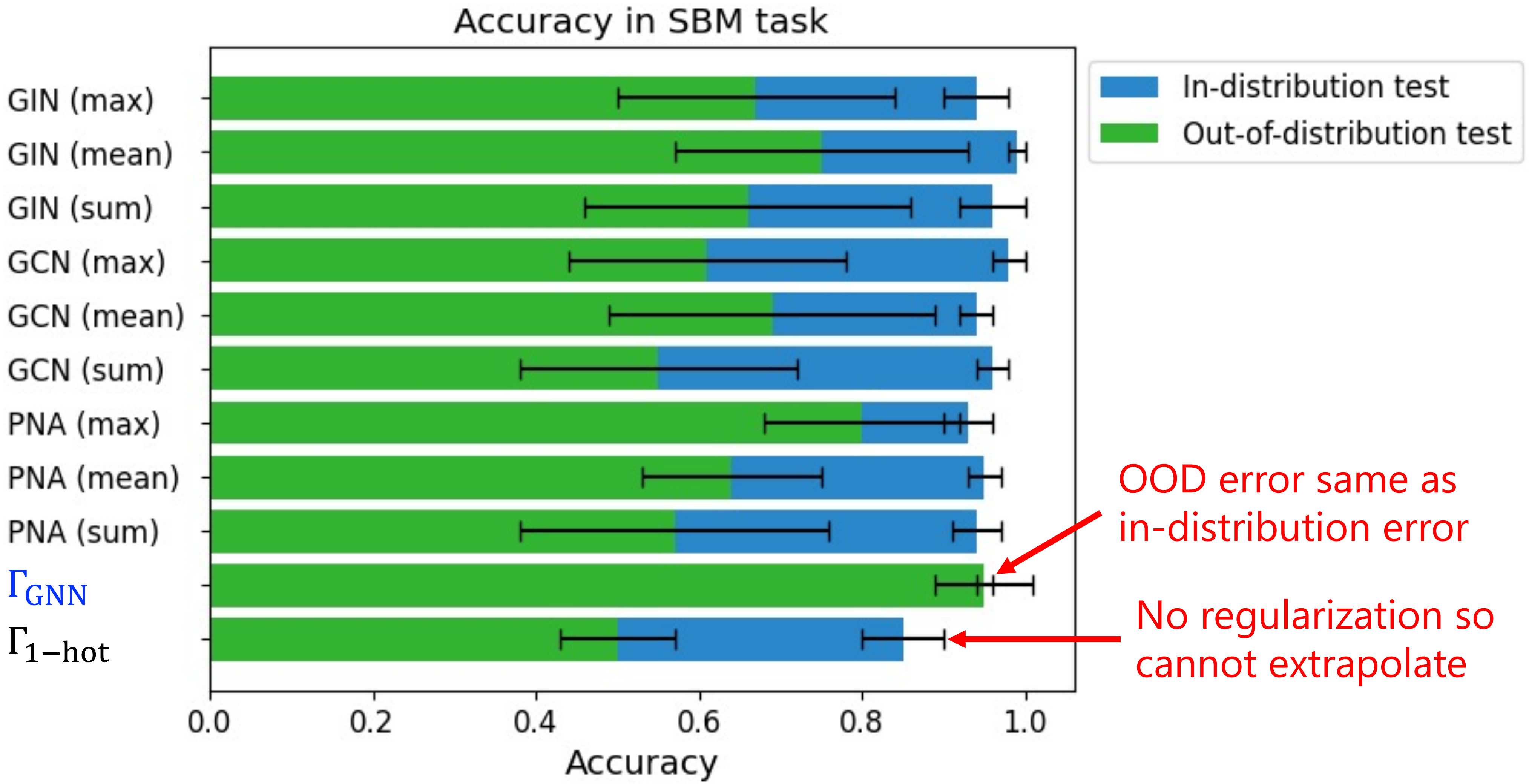
$+ \lambda \ldots$

Pushes subgraph representations towards topology-only unless hurts training loss

Bevilacqua, Zhou, Ribeiro, ICML 2021, Size-Invariant Graph Representations for Graph Classification Extrapolations

OOD Error in Synthetic Task with OOD Attributes

w/ symmetry regularization → Γ_{GNN}



OOD error same as in-distribution error

No regularization so cannot extrapolate

Symmetry regularization helps Γ_{GNN} extrapolate to OOD attributes

Bevilacqua, Zhou, Ribeiro, ICML 2021, Size-Invariant Graph Representations for Graph Classification Extrapolations

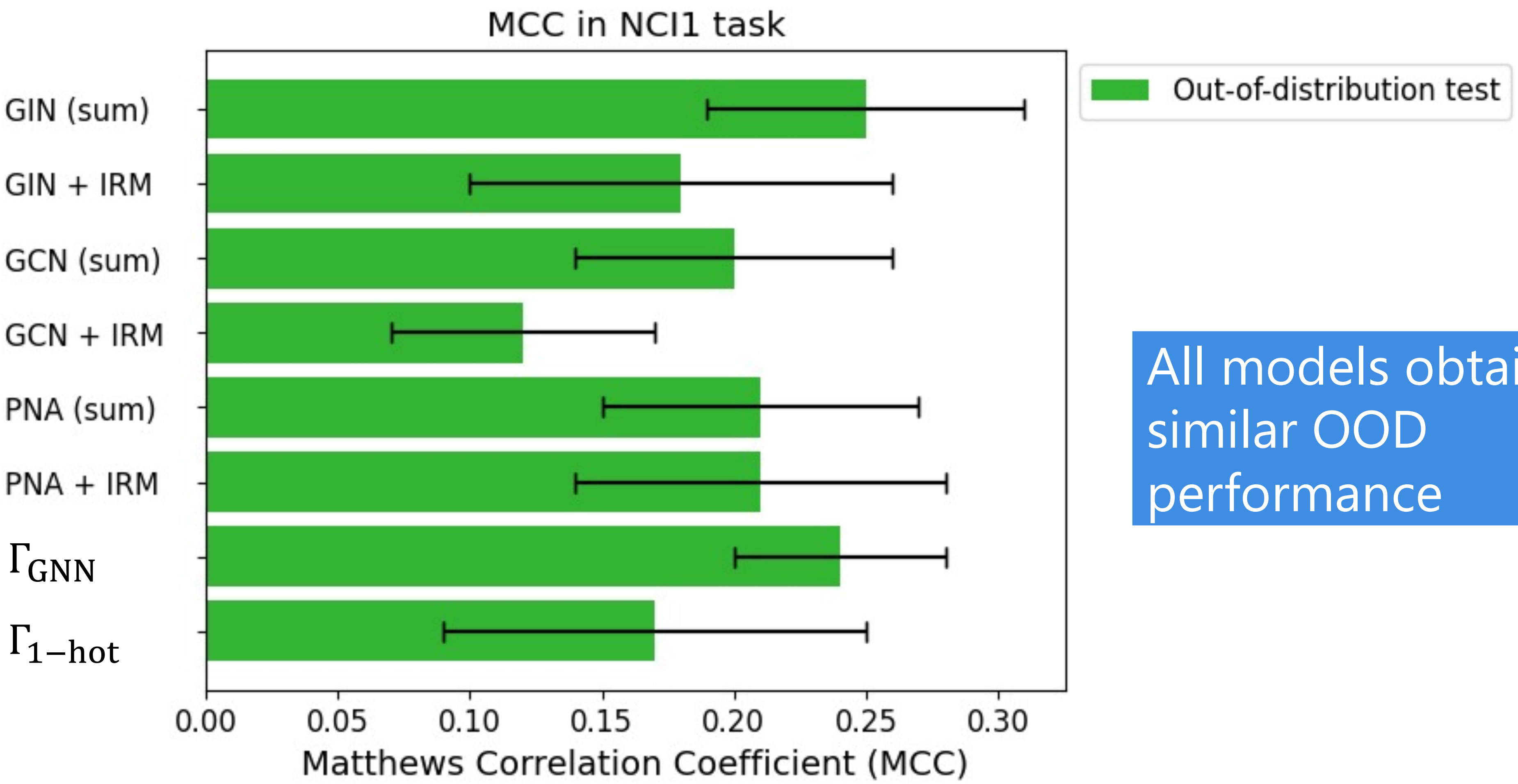
OOD Extrapolation Depends on Causal Mechanism Driving Distribution Shift

I.e.: no OOD universal representations!

No OOD Universal Representations

PURDUE UNIVERSITY

NCI1 task does not follow our causal mechanism



All models obtain similar OOD performance

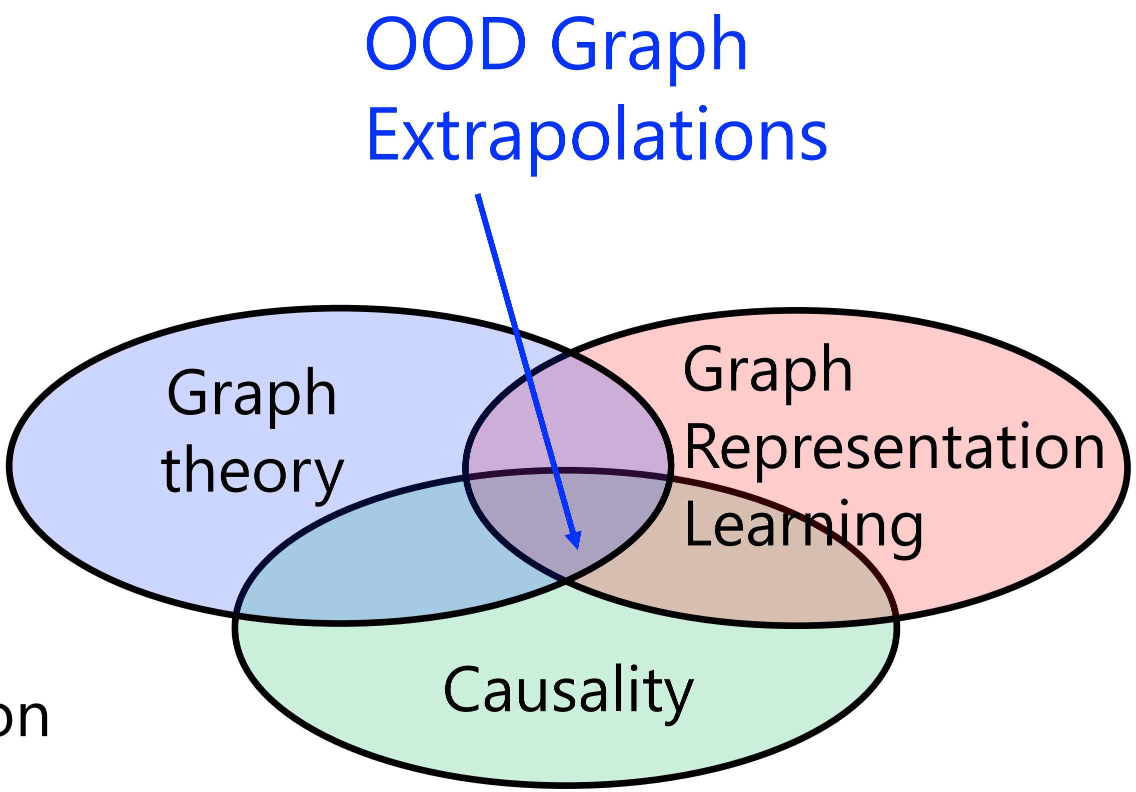
Bevilacqua, Zhou, Ribeiro, ICML 2021, Size-Invariant Graph Representations for Graph Classification Extrapolations

Conclusions

Exciting new area in Graph Representation Learning:

- **OOD extrapolation without examples**
  - Connects counterfactual predictions to stable graph properties
    - E.g., we use subgraph densities as a stable property

- There is no universal OOD graph representation



Thank you!

bbevilac@purdue.edu
zhou950@purdue.edu

References

- Lovász, L. and Szegedy, B. Limits of dense graph sequences. Journal of Combinatorial Theory, Series B, 96(6):933–957, 2006.
- Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019.
- Mouli, S. C. and Ribeiro, B. Neural networks for learning counterfactual g-invariances from single environments. ICLR, 2021.