Learning Causally Invariant Representations for Out-of-Distribution Generalization on Graphs
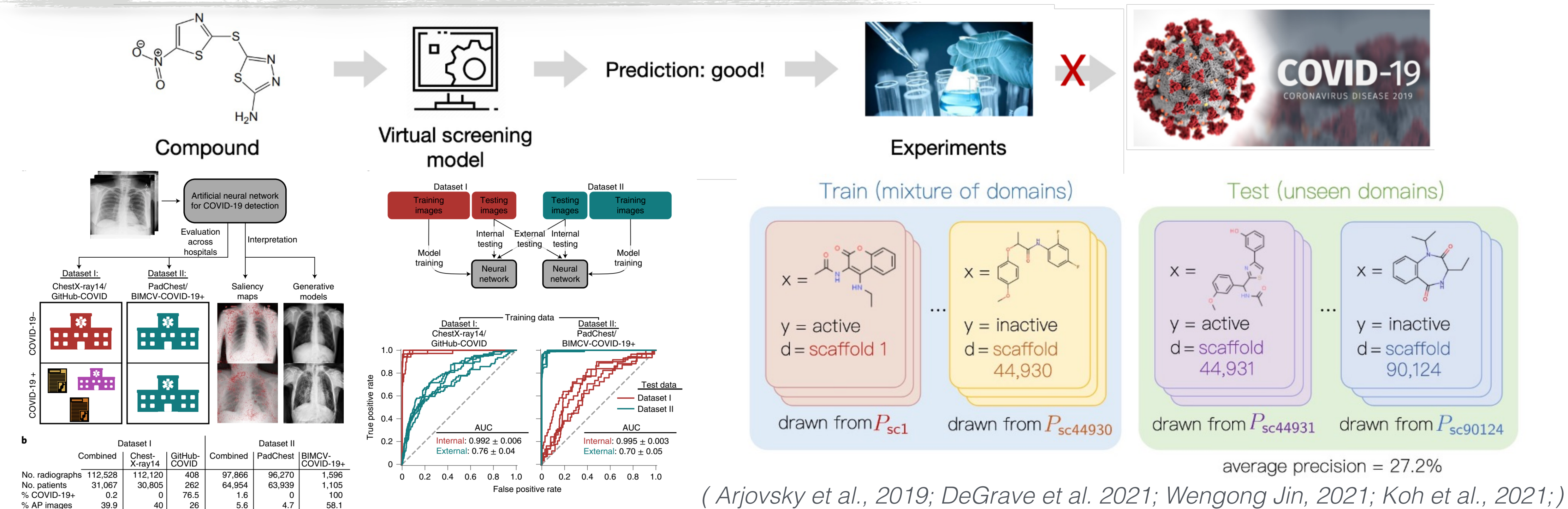
Yongqiang Chen

CUHK, Tencent AI Lab

with Yonggang Zhang, Yatao Bian, Han Yang, Binghui Xie, Kaili Ma, Tongliang Liu, Bo Han, and James Cheng

Out-of-Distribution (OOD) generalization
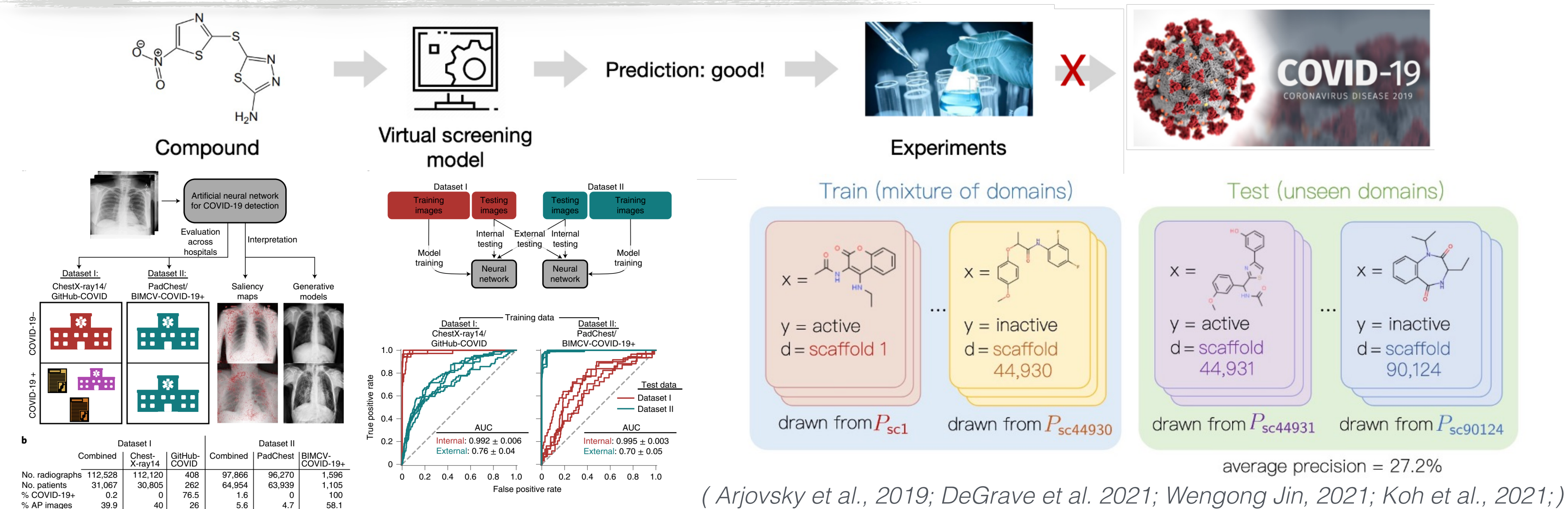

(Arjovsky et al., 2019; DeGrave et al. 2021; Wengong Jin, 2021; Koh et al., 2021;)

Models learned with Empirical Risk Minimization often:

- are prone to **spurious correlations**

- fail catastrophically in **OOD** data
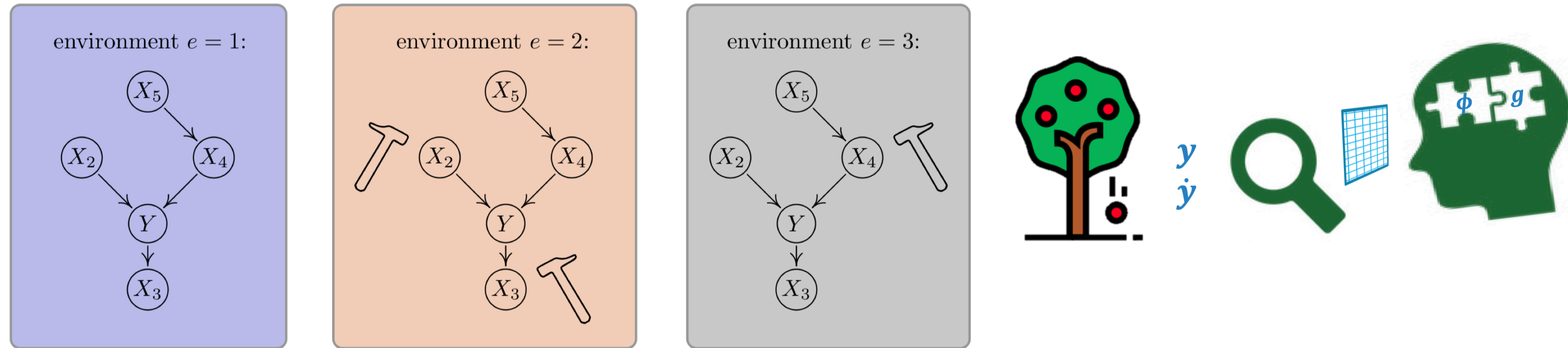
Out-of-Distribution (OOD) generalization

(Arjovsky et al., 2019; DeGrave et al. 2021; Wengong Jin, 2021; Koh et al., 2021;)

The goal of Out-of-Distribution (OOD) generalization:

$$\min_{f:\mathcal{X}\to\mathcal{Y}} \max_{e\in\mathscr{E}_{\text{all}}} \mathscr{L}_e(f)$$

given a subset of training **environments**/domains $\mathscr{E}_{\text{tr}} \subseteq \mathscr{E}_{\text{all}}$,

where each $e \in \mathscr{E}$ corresponds to a dataset $\mathscr{D}_e$ and a loss $\mathscr{L}_e$.
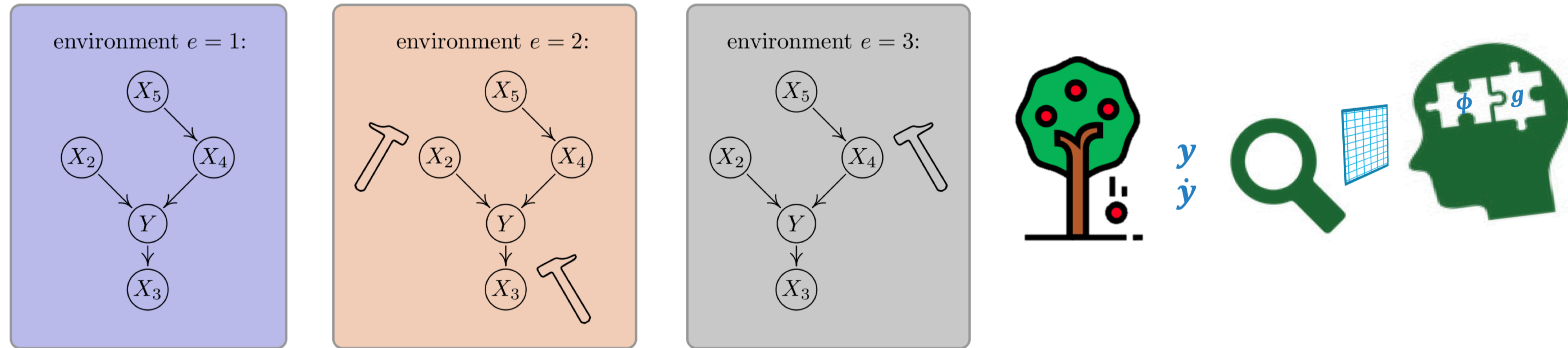
Out-of-Distribution (OOD) generalization

Leveraging the **Invariance Principle** from causality, previous approaches
aim to learn an **invariant** predictor f,

$$\min_{f=w\circ\varphi} \sum_{e\in\mathscr{E}_{\mathrm{tr}}} \mathscr{L}_e(w\circ\varphi),$$

$$\mathrm{s.t.} \ \ w \in \arg\min_{\bar{w}} \mathscr{L}_e(\bar{w}\circ\varphi), \ \forall e \in \mathscr{E}_{\mathrm{tr}},$$

that is **simultaneously optimal** across different environments/domains.

(Peters et al., 2015; Arjovsky et al., 2019; Bottou et al., 2021;)

Out-of-Distribution (OOD) generalization

Previous approaches inspired by the **Invariance Principle** from causality can:

- help to learn the **invariant representations** 😋
- but only works on **linear** regime 🥲
- but only works on **single** distribution shifts 🥲
- but requires **environment**/domain label 🥲

(Peters et al., 2015; Arjovsky et al., 2019; Rosenfeld et al., 2021; Kamath et al., 2021; Ahuja et al., 2021;)

OOD generalization on graphs are more challenging

A Graph Neural Network (GNN) makes predictions taking both **structure-level** and node **attribute-level** features into account.

$$f_{\text{GNN}}\left(\left\{ \text{} \right\}, \left\{ \text{} \right\} \right) = \textbf{``House''}$$
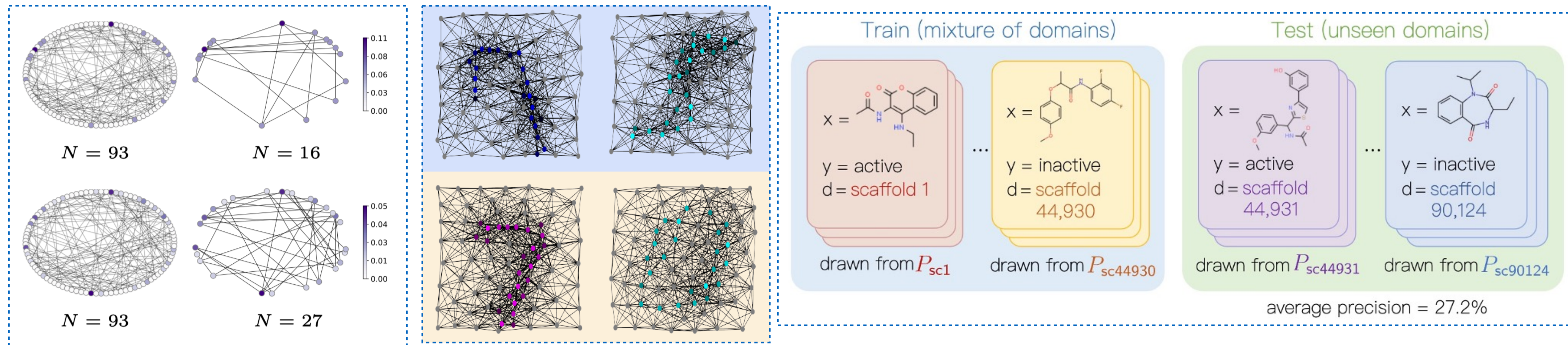
OOD generalization on graphs are more challenging

A Graph Neural Network (GNN) makes predictions taking both **structure-level** and node **attribute-level** features into account.

$$f_{\text{GNN}}\left(\{ \; \} \; , \; \{ \; \} \right) \quad = \quad \text{``House''}$$

Structure-level shifts

Attribute-level shifts

Mixture of structure-level and attribute-level shifts

(Knyazev et al. 2019; Hu et al., 2020; Koh et al., 2021; Gui et al., 2022)

OOD generalization on graphs are more challenging

A Graph Neural Network (GNN) makes predictions taking both **structure-level** and **attribute-level** features into account.

$$f_{\text{GNN}}(\ \{\ \text{⬡}\ \}\ ,\ \{\ \bigcirc\ \bigcirc\ \}\ )\ =\ \text{``House''}$$
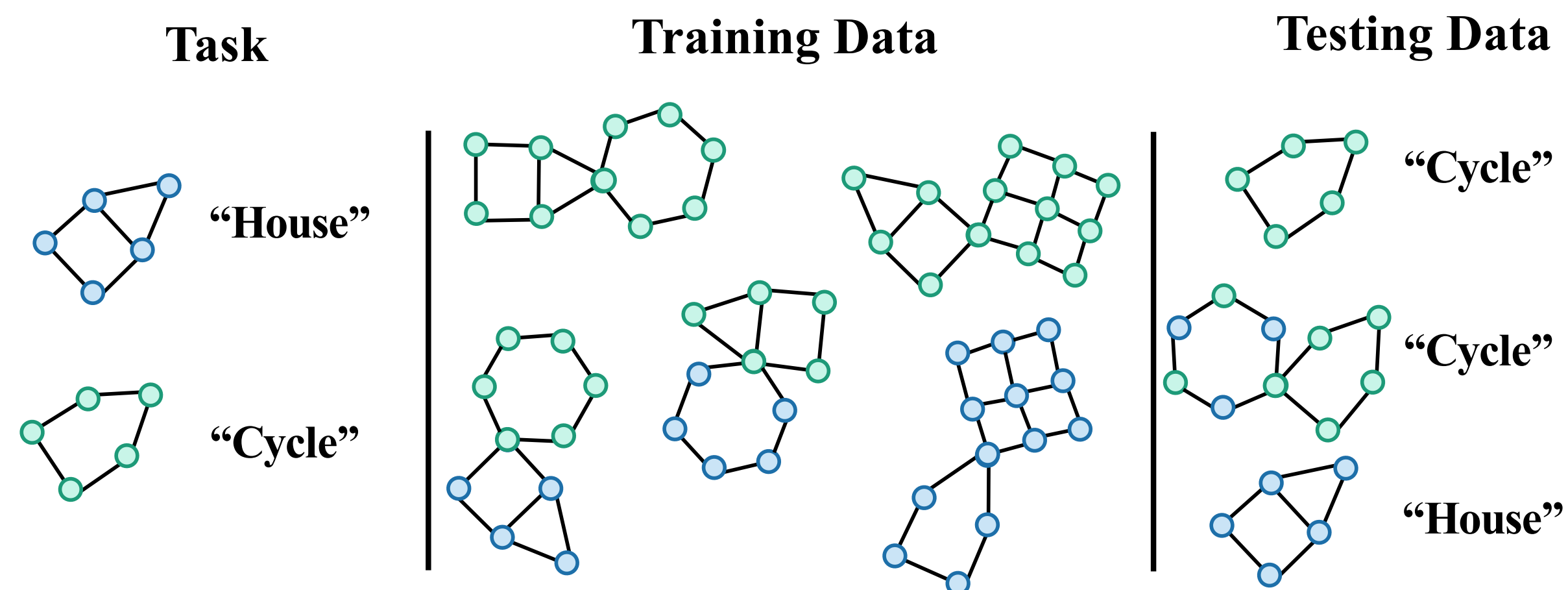

OOD generalization on graphs are **much more challenging!**

(Ying et al., 2019; Luo et al., 2020; Wu et al., 2022;)

OOD generalization on graphs are more challenging

A Graph Neural Network (GNN) makes predictions taking both **structure-level** and **attribute-level** features into account.

$$f_{\text{GNN}}\left(\{ \;\; \} \; , \; \{ \;\; \} \right) \;\; = \;\; \text{``House''}$$

Task   Training Data   Testing Data

"House"

"Cycle"

"Cycle"

"Cycle"

"House"

(Ying et al., 2019; Luo et al., 2020; Wu et al., 2022;)

OOD generalization on graphs are **much more challenging!**

· **Graphs are highly non-linear**

OOD generalization on graphs are more challenging

A Graph Neural Network (GNN) makes predictions taking both **structure-level** and **attribute-level** features into account.

$$f_{\text{GNN}}\left(\ \{\ \text{⬡}\ \}\ ,\ \{\ \bigcirc\ \bigcirc\ \}\ \right)\ =\ \text{“House”}$$

(Ying et al., 2019; Luo et al., 2020; Wu et al., 2022;)

OOD generalization on graphs are **much more challenging!**

- Graphs are highly non-linear
- **Attribute-level shifts**

OOD generalization on graphs are more challenging

A Graph Neural Network (GNN) makes predictions taking both **structure-level** and **attribute-level** features into account.

$$f_{\text{GNN}}\left(\ \{\ \ \ \}\ ,\ \ \{\ \bigcirc\ \bigcirc\ \}\ \right)\ =\ \text{"House"}$$

OOD generalization on graphs are **much more challenging!**
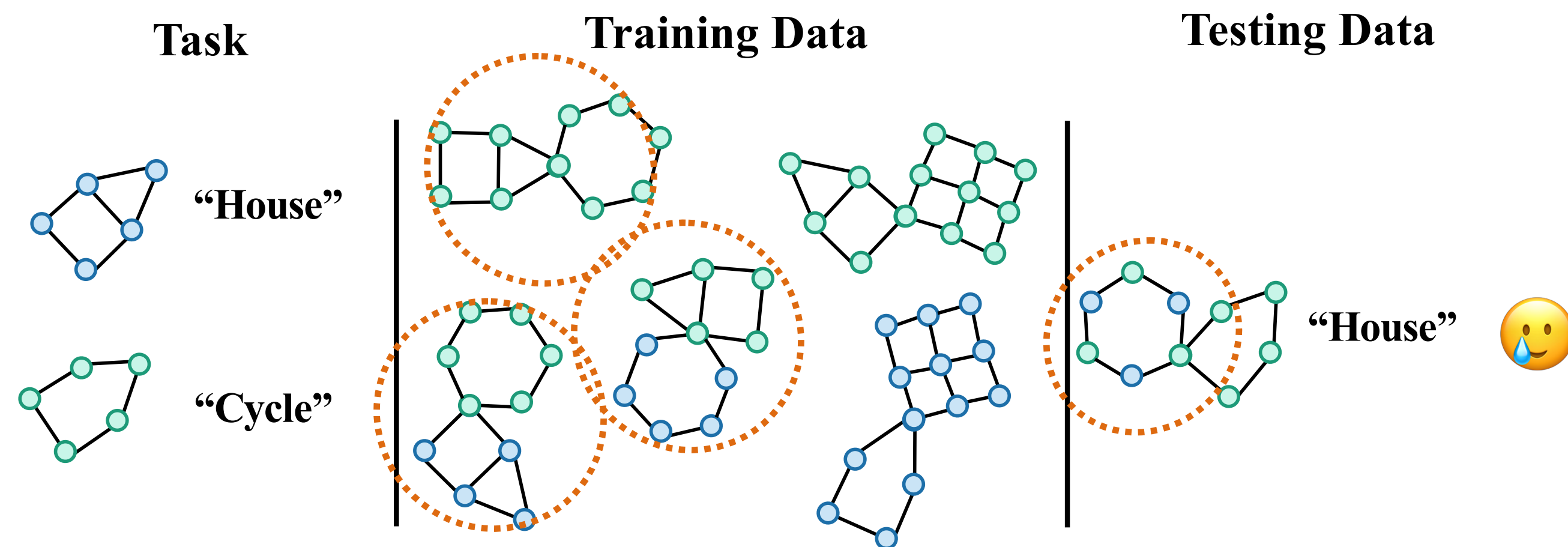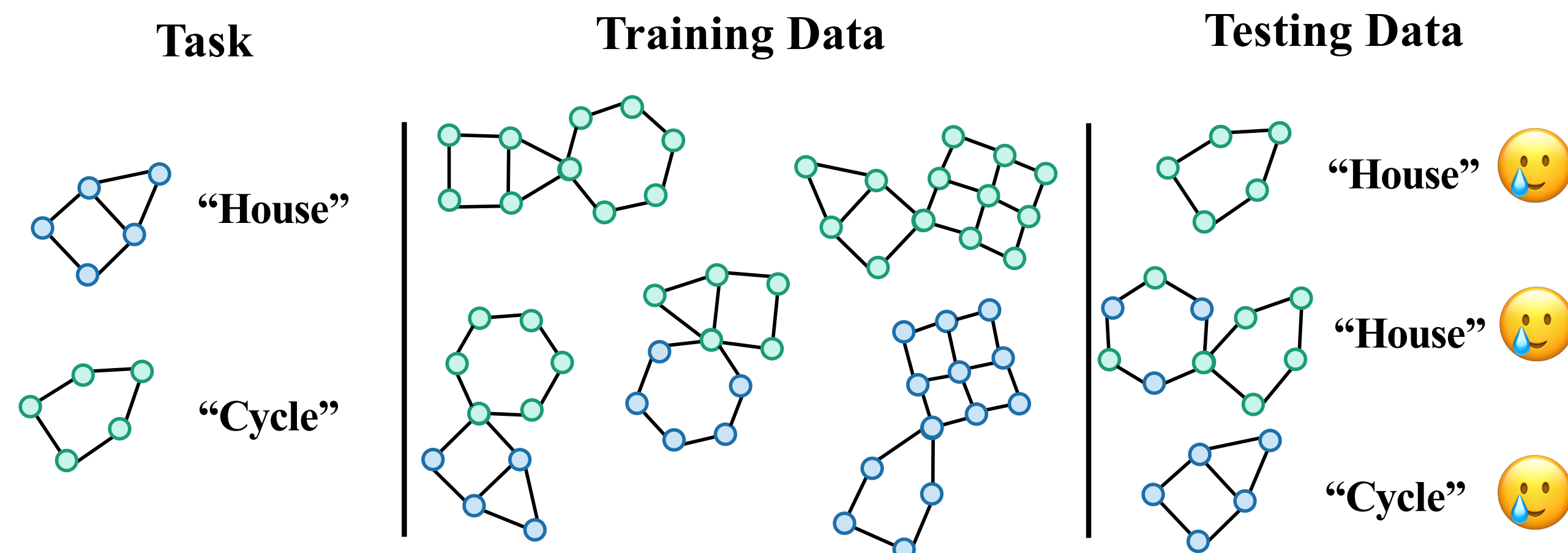
- Graphs are highly non-linear
- Attribute-level shifts
- **Structure-level shifts**

(Ying et al., 2019; Luo et al., 2020; Wu et al., 2022;)

OOD generalization on graphs are more challenging

A Graph Neural Network (GNN) makes predictions taking both **structure-level** and **attribute-level** features into account.

$$f_{\text{GNN}}\left(\{ \; \} \; , \; \{ \; \} \; \right) \quad = \quad \text{"House"}$$

OOD generalization on graphs are **much more challenging!**

- Graphs are highly non-linear
- Attribute-level shifts
- Structure-level shifts
- **Mixed shifts in different modes**

(Ying et al., 2019; Luo et al., 2020; Wu et al., 2022;)

OOD generalization on graphs are more challenging

A Graph Neural Network (GNN) makes predictions taking both **structure-level** and **attribute-level** features into account.

$$f_{\text{GNN}}\left(\ \{\ \text{⬡} \ \}\ ,\ \{\ \bigcirc\ \bigcirc\ \}\ \right)\quad =\quad \text{"House"}$$

No environment partitions

Task

"House"

"Cycle"

Training Data

Testing Data

"House" 😢

"House" 😢

"Cycle" 😢

(Ying et al., 2019; Luo et al., 2020; Wu et al., 2022;)

OOD generalization on graphs are **much more challenging!**

- Graphs are highly non-linear
- Attribute-level shifts
- Structure-level shifts
- Mixed shifts in different modes
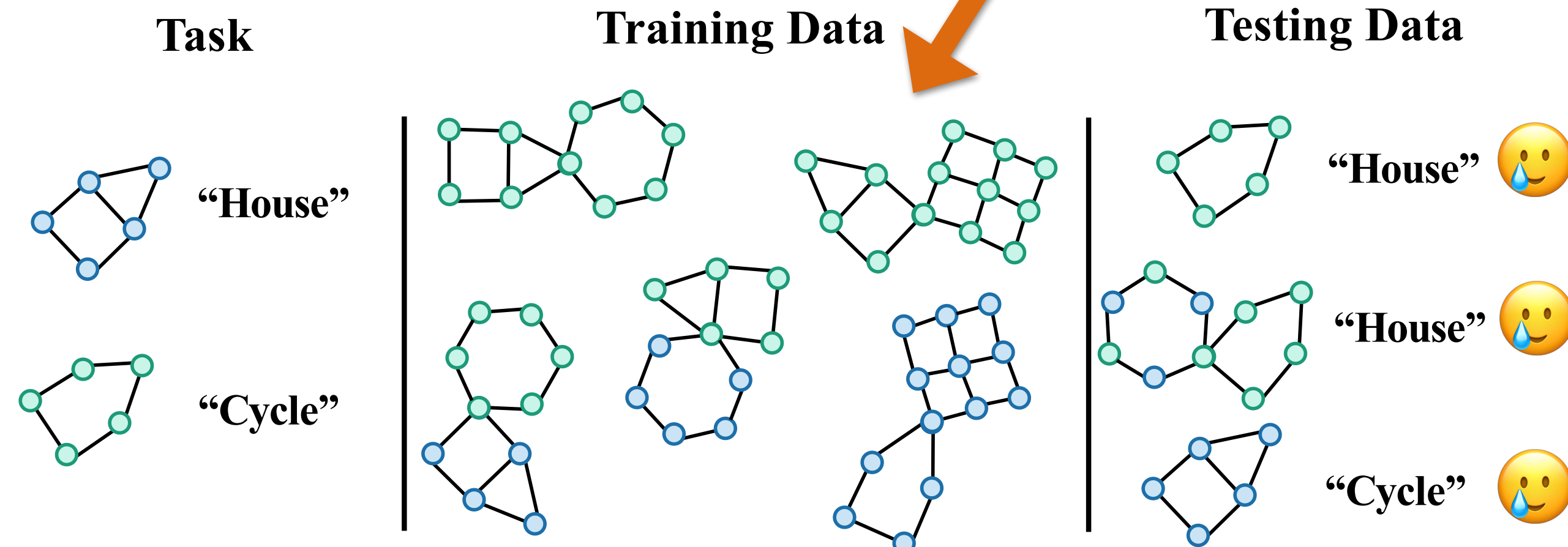- **Expensive environment labels**

OOD generalization on graphs are more challenging

Task

Training Data

Testing Data

"House"

"Cycle"

"House" 😢

"House" 😢

"Cycle" 😢

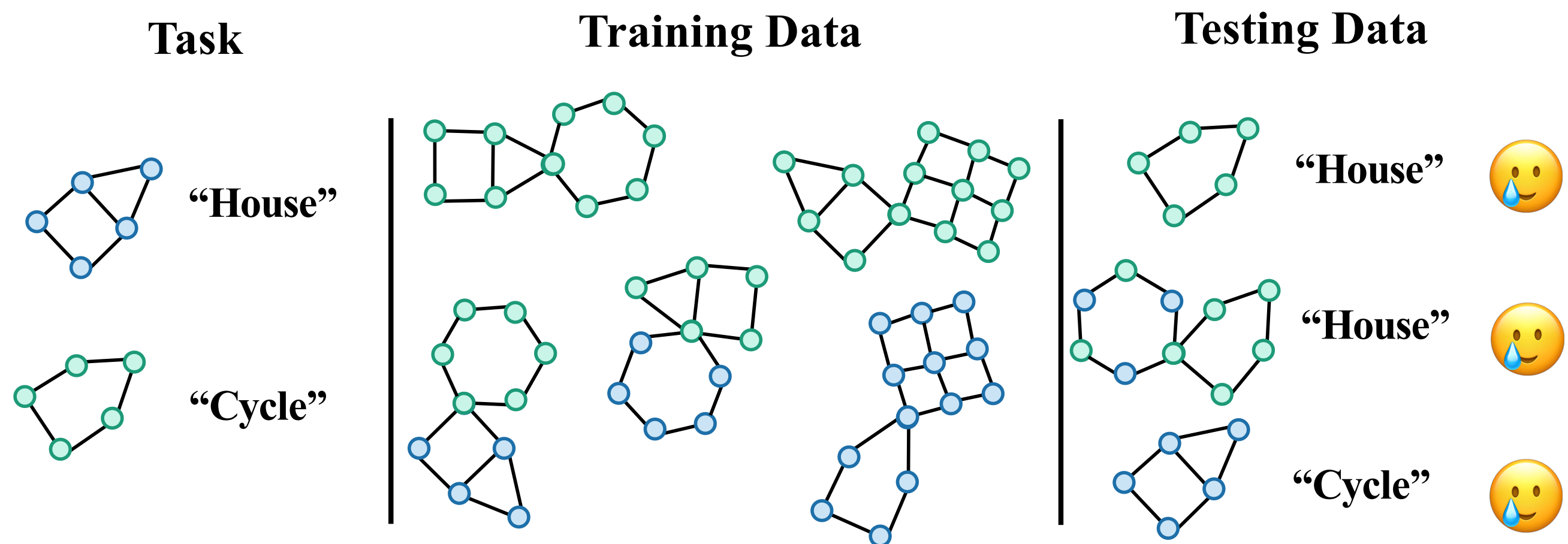*(Ying et al., 2019; Luo et al., 2020; Wu et al., 2022;)*

OOD generalization on graphs are **much more challenging!**

- Graphs are highly non-linear
- Attribute-level shifts
- Structure-level shifts
- Mixed shifts in different modes
- Expensive environment labels

Structure and attribute shifts

Mixed with **graph size** shifts

Structure and attribute shifts

OOD failures of GNNs *training objectives* and *architectures*

14

OOD generalization on graphs are more challenging

Task

Training Data

Testing Data
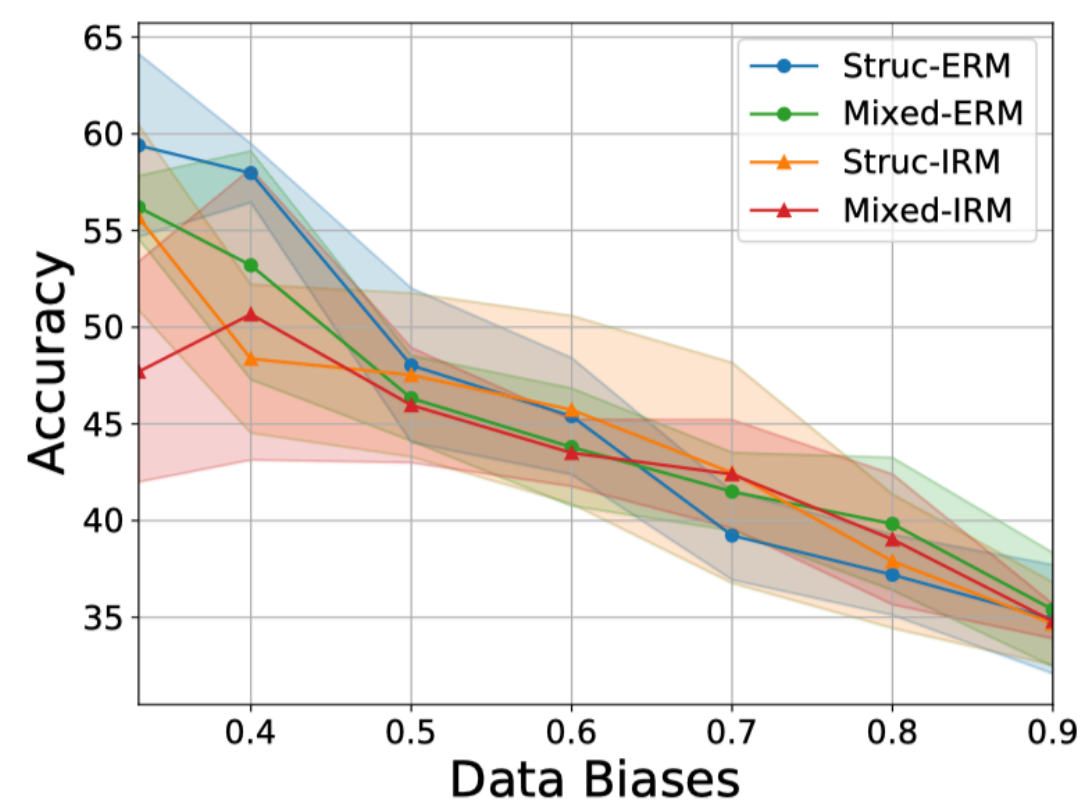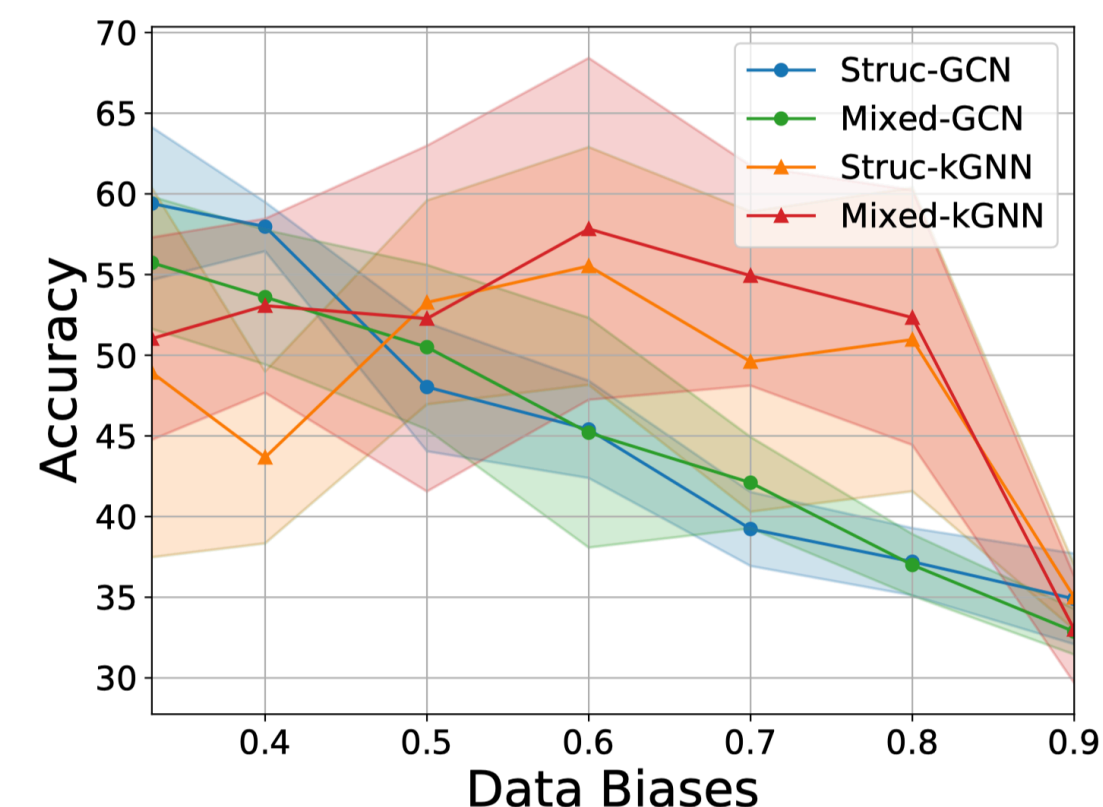
"House"

"House" 🥲

"Cycle"

OOD generalization on graphs are **much more challenging!**

- Graphs are highly non-linear

modes

As existing approaches are down…
How can we define and capture the invariance on graphs?
Can we train a GNN that is generalizable to OOD graphs?

Structure and attribute shifts

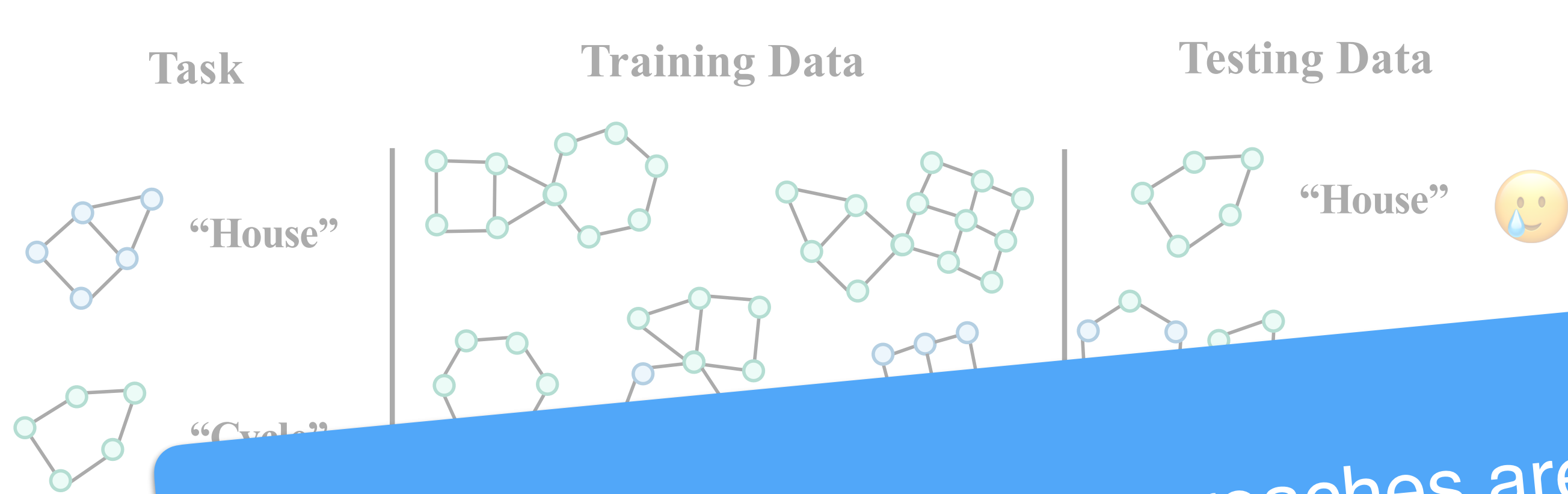Mixed with **graph size** shifts

Structure and attribute shifts

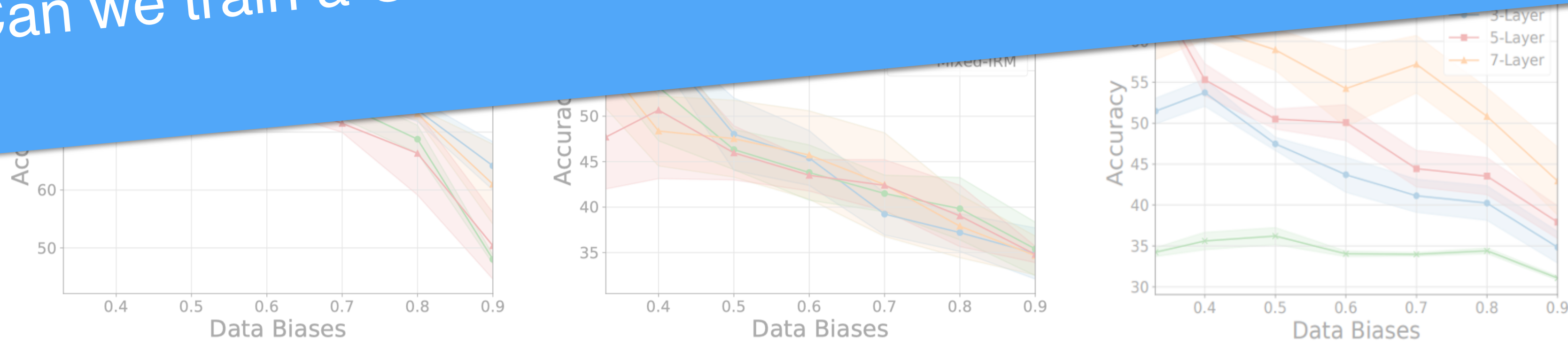OOD failures of GNNs *training objectives* and *architectures*

Invariance Principle Meets Graph Neural Networks

for generalizing to out-of-distribution graph data

Figure source: Léon Bottou

Structural Causal Models for graph generalization

Graph Generation Process:

$$f_{\text{gen}} : \mathscr{Z} \to \mathscr{G}$$

C

Invariant features

S

Spurious features

(a) $\mathcal{G}$-Gen. SCM   (b) FIIF SCM   (c) PIIF SCM

Structural Causal Models

Structural Causal Models for graph generalization

Graph Generation Process:

$$f_{\text{gen}} : \mathcal{Z} \to \mathcal{G}$$

C

S

Invariant features

Spurious features

Structural Causal Models

(a) $\mathcal{G}$-Gen. SCM (b) FIIF SCM (c) PIIF SCM

Soluble or not?

Realistic examples

(Yu et al., 2021; Miao et al., 2022;)

CIGA: Causality Inspired Invariant Graph LeArning

Step 1: Invariant subgraph identification

Featurizer GNN $g : \mathcal{G} \rightarrow \mathcal{G}_c$

Structural Causal Models

CIGA: Causality Inspired Invariant Graph LeArning

Step 1: Invariant subgraph identification

Featurizer GNN $g : \mathcal{G} \to \mathcal{G}_c$

(a) $\mathcal{G}$-Gen. SCM   (b) FIIF SCM   (c) PIIF SCM

Structural Causal Models

Step 2: Label prediction

Classifier GNN $f_c : \mathcal{G}_c \to \mathcal{Y}$

Invariant Subgraph Identification   Classification

g f_c "House"   "Cycle"

$\widehat{G}_c$

House

Cycle

CIGA: Causality Inspired Invariant Graph LeArning

Step 1: Invariant subgraph identification

Featurizer GNN $g : \mathscr{G} \to \mathscr{G}_c$

Structural Causal Models

(a) $\mathcal{G}$-Gen. SCM   (b) FIIF SCM   (c) PIIF SCM

Invariant Subgraph Identification   Classification

g f_c "House"   "Cycle"

$\widehat{G}_c$
House
Cycle

Step 2: Label prediction

Classifier GNN $f_c : \mathscr{G}_c \to \mathscr{Y}$

Overall objective

$$\max_{f_c, g} I(\widehat{G}_c; Y), \text{ s.t. } \widehat{G}_c \perp\!\!\!\perp E, \ \widehat{G}_c = g(G),$$

Informative   Invariant

CIGA: Causality Inspired Invariant Graph LeArning

CIGAv1: when $|G_c| = s_c$ is known and fixed

$$\max_{f_c,g} I(\widehat{G}_c;Y), \text{ s.t. } \widehat{G}_c \in \underset{\widehat{G}_c=g(G),|\widehat{G}_c|\leq s_c}{\arg\max} I(\widehat{G}_c;\widetilde{G}_c|Y),$$

(a) $\mathcal{G}$-Gen. SCM   (b) FIIF SCM   (c) PIIF SCM

Structural Causal Models

Invariant Subgraph Identification   Classification

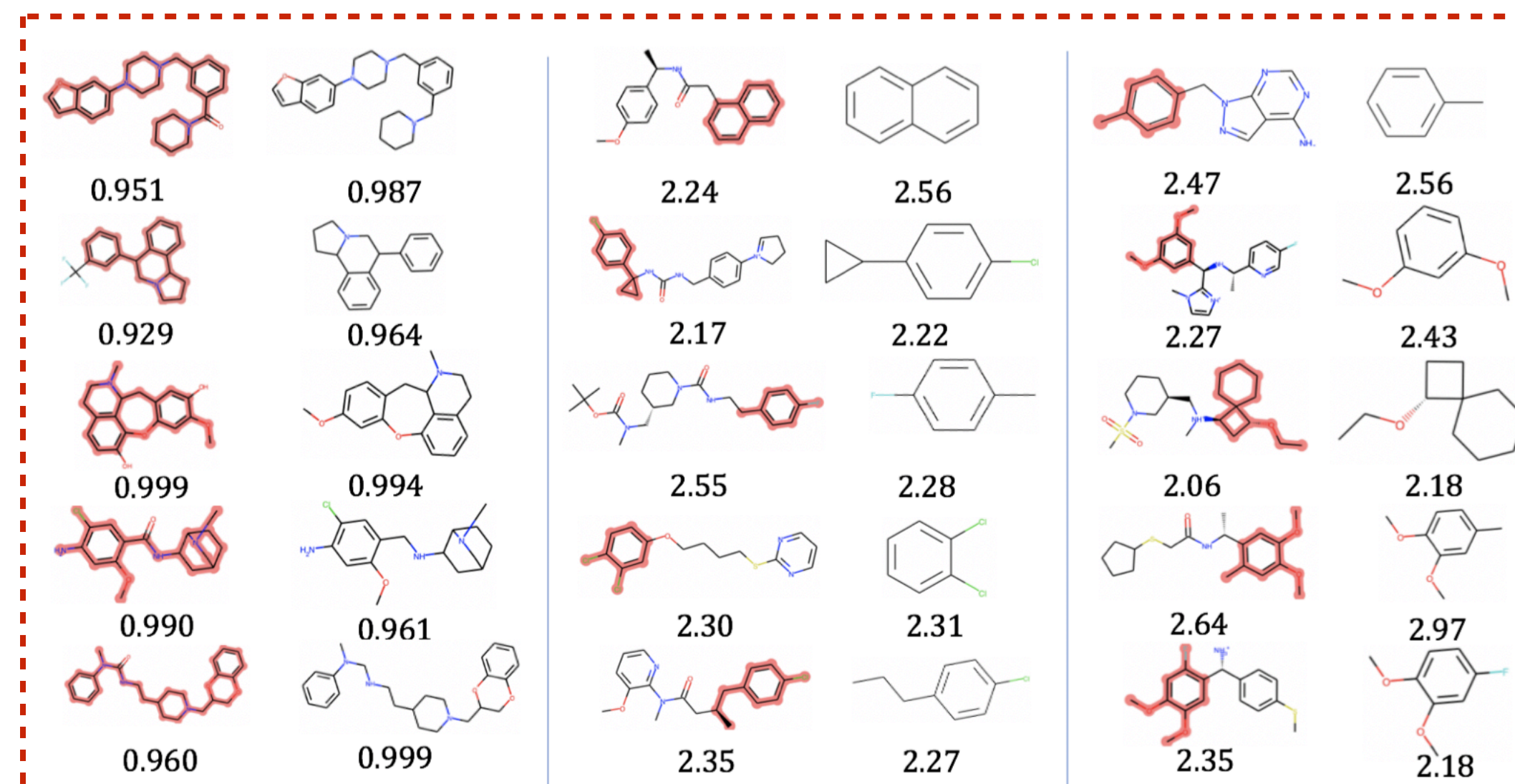CIGAv2: eliminate the size constraint

$$\max_{f_c,g} I(\widehat{G}_c;Y) + I(\widehat{G}_s;Y), \text{ s.t. } \widehat{G}_c \in \arg\max_{\widehat{G}_c=g(G)} I(\widehat{G}_c;\widetilde{G}_c|Y),$$
$$I(\widehat{G}_s;Y) \leq I(\widehat{G}_c;Y), \widehat{G}_s = G - g(G),$$

22

Experiments on synthetic graph distribution shifts

Theoretical results (Informal):
Given the previous SCMs, each solution to CIGAv1 or CIGAv2 elicits a GNN that is generalizable against various distribution shifts, with some mild assumptions on training environments, and the expressivity of GNNs encoders.

Table 1: OOD generalization performance on structure and mixed shifts for synthetic graphs.

	SPMotif-Struc†			SPMotif-Mixed†			AVG
	BIAS=0.33	BIAS=0.60	BIAS=0.90	BIAS=0.33	BIAS=0.60	BIAS=0.90	
ERM	59.49 (3.50)	55.48 (4.84)	49.64 (4.63)	58.18 (4.30)	49.29 (8.17)	41.36 (3.29)	52.24
ASAP	64.87 (13.8)	64.85 (10.6)	**57.29 (14.5)**	66.88 (15.0)	59.78 (6.78)	**50.45 (4.90)**	60.69
DIR	58.73 (11.9)	48.72 (14.8)	41.90 (9.39)	67.28 (4.06)	51.66 (14.1)	38.58 (5.88)	51.14
IRM	57.15 (3.98)	61.74 (1.32)	45.68 (4.88)	58.20 (1.97)	49.29 (3.67)	40.73 (1.93)	52.13
V-REX	54.64 (3.05)	53.60 (3.74)	48.86 (9.69)	57.82 (5.93)	48.25 (2.79)	43.27 (1.32)	51.07
EIIL	56.48 (2.56)	60.07 (4.47)	55.79 (6.54)	53.91 (3.15)	48.41 (5.53)	41.75 (4.97)	52.73
IB-IRM	58.30 (6.37)	54.37 (7.35)	45.14 (4.07)	57.70 (2.11)	50.83 (1.51)	40.27 (3.68)	51.10
CNC	70.44 (2.55)	**66.79 (9.42)**	50.25 (10.7)	65.75 (4.35)	59.27 (5.29)	41.58 (1.90)	59.01
CIGAv1	**71.07 (3.60)**	63.23 (9.61)	51.78 (7.29)	**74.35 (1.85)**	**64.54 (8.19)**	49.01 (9.92)	**62.33**
CIGAv2	**77.33 (9.13)**	**69.29 (3.06)**	**63.41 (7.38)**	**72.42 (4.80)**	**70.83 (7.54)**	**54.25 (5.38)**	**67.92**
ORACLE (IID)		88.70 (0.17)			88.73 (0.25)		

†Higher accuracy and lower variance indicate better OOD generalization ability.

CIGA outperforms previous methods under ***structure and mixed shifts*** by a significant margin up to ***10%***.

Experiments on realistic graph distribution shifts

Theoretical results (Informal):
Given the previous SCMs, each solution to CIGAv1 or CIGAv2 elicits a GNN that is generalizable against various distribution shifts, with some mild assumptions on training environments, and the expressivity of GNNs encoders.

Table 2: OOD generalization performance on complex distribution shifts for real-world graphs.

DATASETS	DRUG-ASSAY	DRUG-SCA	DRUG-SIZE	CMNIST-SP	GRAPH-SST5	TWITTER	AVG (RANK)[†]
ERM	71.79 (0.27)	68.85 (0.62)	66.70 (1.08)	13.96 (5.48)	43.89 (1.73)	60.81 (2.05)	54.33 (6.00)
ASAP	70.51 (1.93)	66.19 (0.94)	64.12 (0.67)	10.23 (0.51)	44.16 (1.36)	60.68 (2.10)	52.65 (8.33)
GIB	63.01 (1.16)	62.01 (1.41)	55.50 (1.42)	15.40 (3.91)	38.64 (4.52)	48.08 (2.27)	47.11 (10.0)
DIR	68.25 (1.40)	63.91 (1.36)	60.40 (1.42)	15.50 (8.65)	41.12 (1.96)	59.85 (2.98)	51.51 (9.33)
IRM	72.12 (0.49)	68.69 (0.65)	66.54 (0.42)	31.58 (9.52)	43.69 (1.26)	63.50 (1.23)	57.69 (4.50)
V-REX	72.05 (1.25)	68.92 (0.98)	66.33 (0.74)	10.29 (0.46)	43.28 (0.52)	63.21 (1.57)	54.01 (6.17)
EIIL	72.60 (0.47)	68.45 (0.53)	66.38 (0.66)	30.04 (10.9)	42.98 (1.03)	62.76 (1.72)	57.20 (5.33)
IB-IRM	72.50 (0.49)	68.50 (0.40)	66.64 (0.28)	**39.86 (10.5)**	40.85 (2.08)	61.26 (1.20)	58.27 (5.33)
CNC	72.40 (0.46)	67.24 (0.90)	65.79 (0.80)	12.21 (3.85)	42.78 (1.53)	61.03 (2.49)	53.56 (7.50)
CIGAv1	**72.71 (0.52)**	**69.04 (0.86)**	**67.24 (0.88)**	19.77 (17.1)	**44.71 (1.14)**	**63.66 (0.84)**	**56.19 (2.50)**
CIGAv2	**73.17 (0.39)**	**69.70 (0.27)**	**67.78 (0.76)**	**44.91 (4.31)**	**45.25 (1.27)**	**64.45 (1.99)**	**60.88 (1.00)**
ORACLE (IID)	85.56 (1.44)	84.71 (1.60)	85.83 (1.31)	62.13 (0.43)	48.18 (1.00)	64.21 (1.77)	

[†]Averaged rank is also reported in the blankets because of dataset heterogeneity. Lower rank is better.

Table 3: OOD generalization performance on graph size shifts for real-world graphs in terms of Matthews correlation coefficient.

DATASETS	NCI1	NCI109	PROTEINS	DD	AVG
ERM	0.15 (0.05)	0.16 (0.02)	0.22 (0.09)	0.27 (0.09)	0.20
ASAP	0.16 (0.10)	0.15 (0.07)	0.22 (0.16)	0.21 (0.08)	0.19
GIB	0.13 (0.10)	0.16 (0.02)	0.19 (0.08)	0.01 (0.18)	0.12
DIR	0.21 (0.06)	0.13 (0.05)	0.25 (0.14)	0.20 (0.10)	0.20
IRM	0.17 (0.02)	0.14 (0.01)	0.21 (0.09)	0.22 (0.08)	0.19
V-REX	0.15 (0.04)	0.15 (0.04)	0.22 (0.06)	0.21 (0.07)	0.18
EIIL	0.14 (0.03)	0.16 (0.02)	0.20 (0.05)	0.23 (0.10)	0.19
IB-IRM	0.12 (0.04)	0.15 (0.06)	0.21 (0.06)	0.15 (0.13)	0.16
CNC	0.16 (0.04)	0.16 (0.04)	0.19 (0.08)	0.27 (0.13)	0.20
WL KERNEL	**0.39 (0.00)**	0.21 (0.00)	0.00 (0.00)	0.00 (0.00)	0.15
GC KERNEL	0.02 (0.00)	0.00 (0.00)	0.29 (0.00)	0.00 (0.00)	0.08
$\Gamma_{1\text{-HOT}}$	0.17 (0.08)	**0.25 (0.06)**	0.12 (0.09)	0.23 (0.08)	0.19
Γ_{GIN}	0.24 (0.04)	0.18 (0.04)	0.29 (0.11)	**0.28 (0.06)**	0.25
Γ_{RPGIN}	0.26 (0.05)	0.20 (0.04)	0.25 (0.12)	0.20 (0.05)	0.23
CIGAv1	0.22 (0.07)	**0.23 (0.09)**	**0.40 (0.06)**	**0.29 (0.08)**	**0.29**
CIGAv2	**0.27 (0.07)**	0.22 (0.05)	0.31 (0.12)	0.26 (0.08)	**0.27**
ORACLE (IID)	0.32 (0.05)	0.37 (0.06)	0.39 (0.09)	0.33 (0.05)	

CIGA outperforms previous methods under other ***realistic shifts*** by a significant margin up to ***10%***.

Experiments on realistic graph distribution shifts

> CIGA is the *only* approach that *consistently* outperforms ERM in the challenging setting from AI-aided drug discovery — **DrugOOD**.

Table 2: OOD generalization performance on complex distribution shifts for real-world graphs.

DATASETS	DRUG-ASSAY	DRUG-SCA	DRUG-SIZE	CMNIST-SP	GRAPH-SST5	TWITTER	AVG (RANK)[†]
ERM	71.79 (0.27)	68.85 (0.62)	66.70 (1.08)	13.96 (5.48)	43.89 (1.73)	60.81 (2.05)	54.33 (6.00)
ASAP	70.51 (1.93)	66.19 (0.94)	64.12 (0.67)	10.23 (0.51)	44.16 (1.36)	60.68 (2.10)	52.65 (8.33)
GIB	63.01 (1.16)	62.01 (1.41)	55.50 (1.42)	15.40 (3.91)	38.64 (4.52)	48.08 (2.27)	47.11 (10.0)
DIR	68.25 (1.40)	63.91 (1.36)	60.40 (1.42)	15.50 (8.65)	41.12 (1.96)	59.85 (2.98)	51.51 (9.33)
IRM	72.12 (0.49)	68.69 (0.65)	66.54 (0.42)	31.58 (9.52)	43.69 (1.26)	63.50 (1.23)	57.69 (4.50)
V-REX	72.05 (1.25)	68.92 (0.98)	66.33 (0.74)	10.29 (0.46)	43.28 (0.52)	63.21 (1.57)	54.01 (6.17)
EIIL	72.60 (0.47)	68.45 (0.53)	66.38 (0.66)	30.04 (10.9)	42.98 (1.03)	62.76 (1.72)	57.20 (5.33)
IB-IRM	72.50 (0.49)	68.50 (0.40)	66.64 (0.28)	**39.86 (10.5)**	40.85 (2.08)	61.26 (1.20)	58.27 (5.33)
CNC	72.40 (0.46)	67.24 (0.90)	65.79 (0.80)	12.21 (3.85)	42.78 (1.53)	61.03 (2.49)	53.56 (7.50)
CIGAv1	**72.71 (0.52)**	**69.04 (0.86)**	**67.24 (0.88)**	19.77 (17.1)	**44.71 (1.14)**	**63.66 (0.84)**	**56.19 (2.50)**
CIGAv2	**73.17 (0.39)**	**69.70 (0.27)**	**67.78 (0.76)**	44.91 (4.31)	45.25 (1.27)	64.45 (1.99)	**60.88 (1.00)**
ORACLE (IID)	85.56 (1.44)	84.71 (1.60)	85.83 (1.31)	62.13 (0.43)	48.18 (1.00)	64.21 (1.77)	

[†]Averaged rank is also reported in the blankets because of dataset heterogeneity. Lower rank is better.
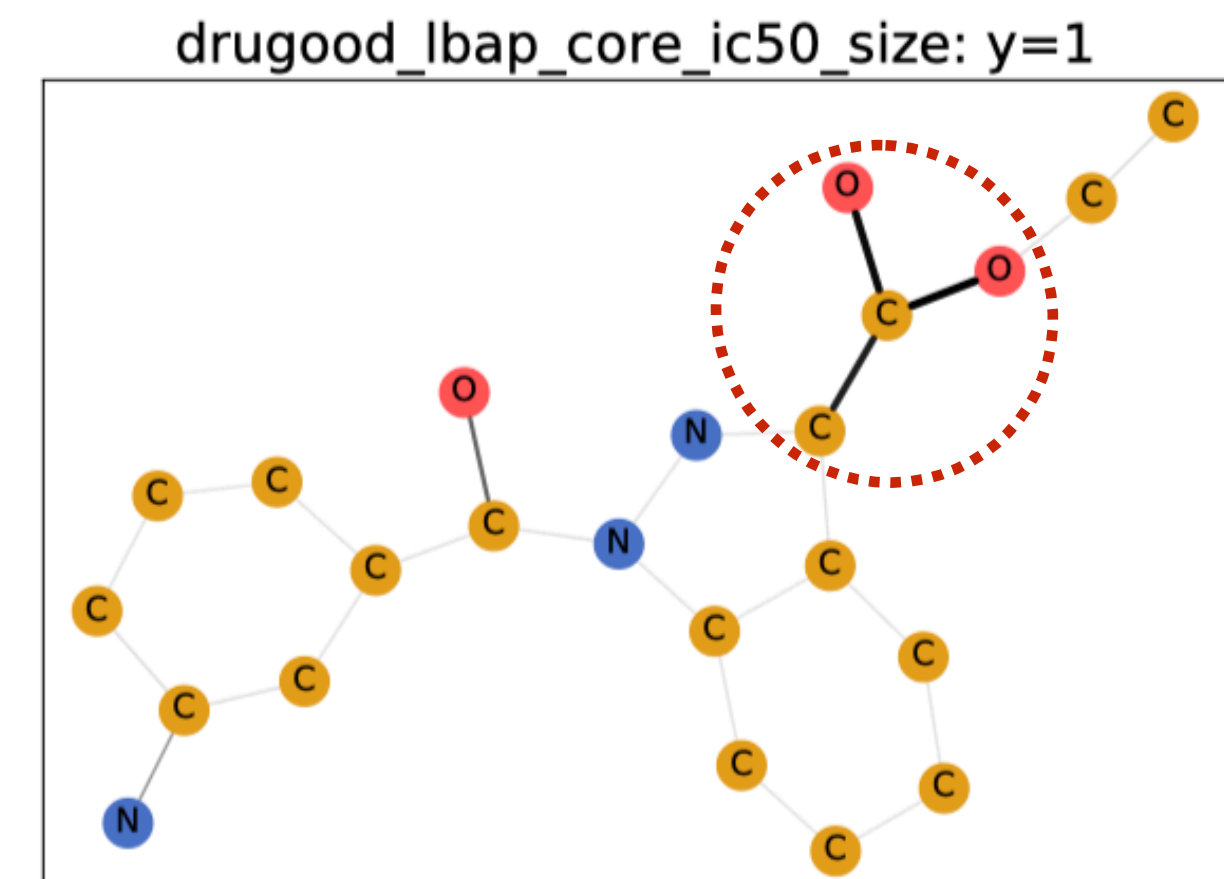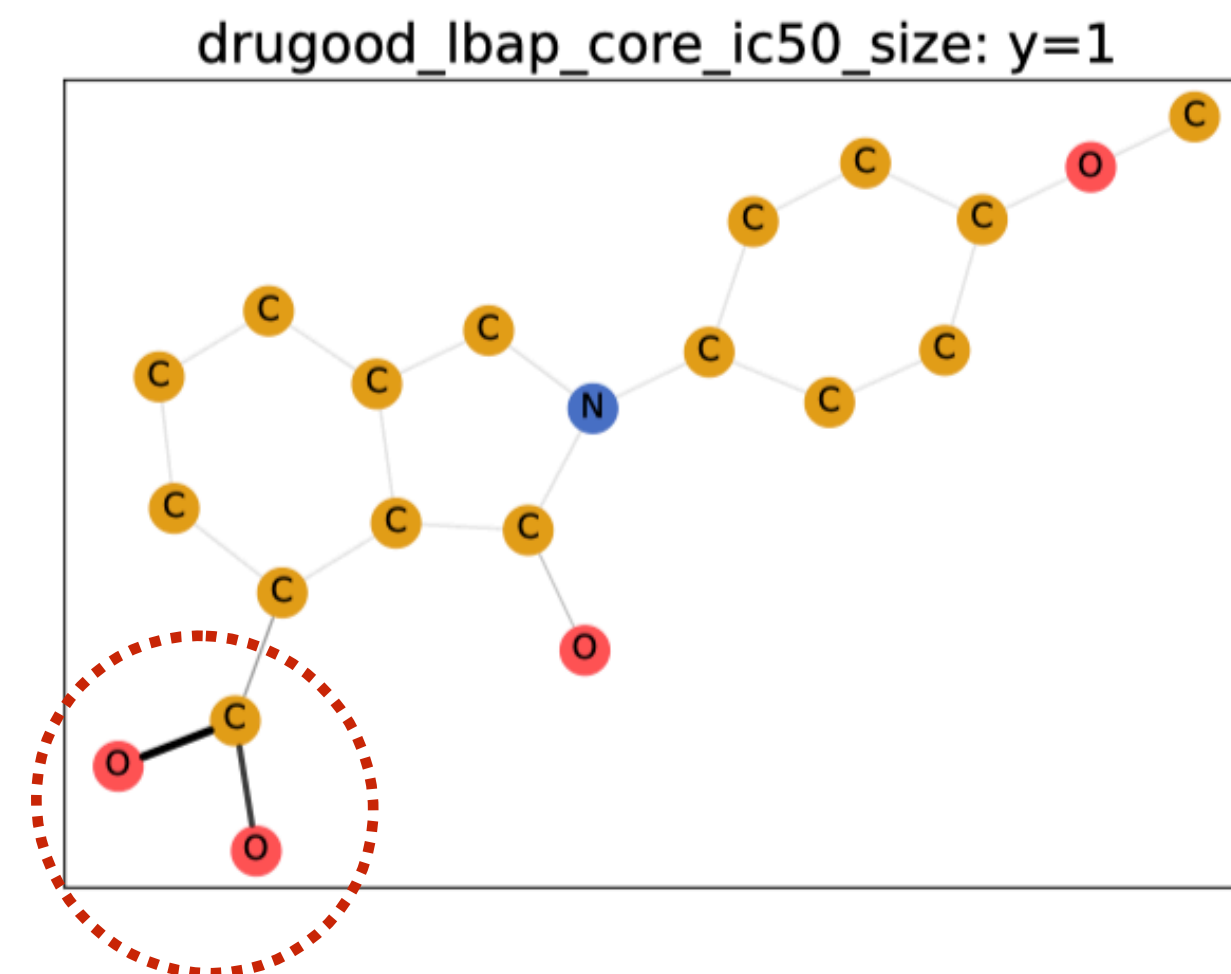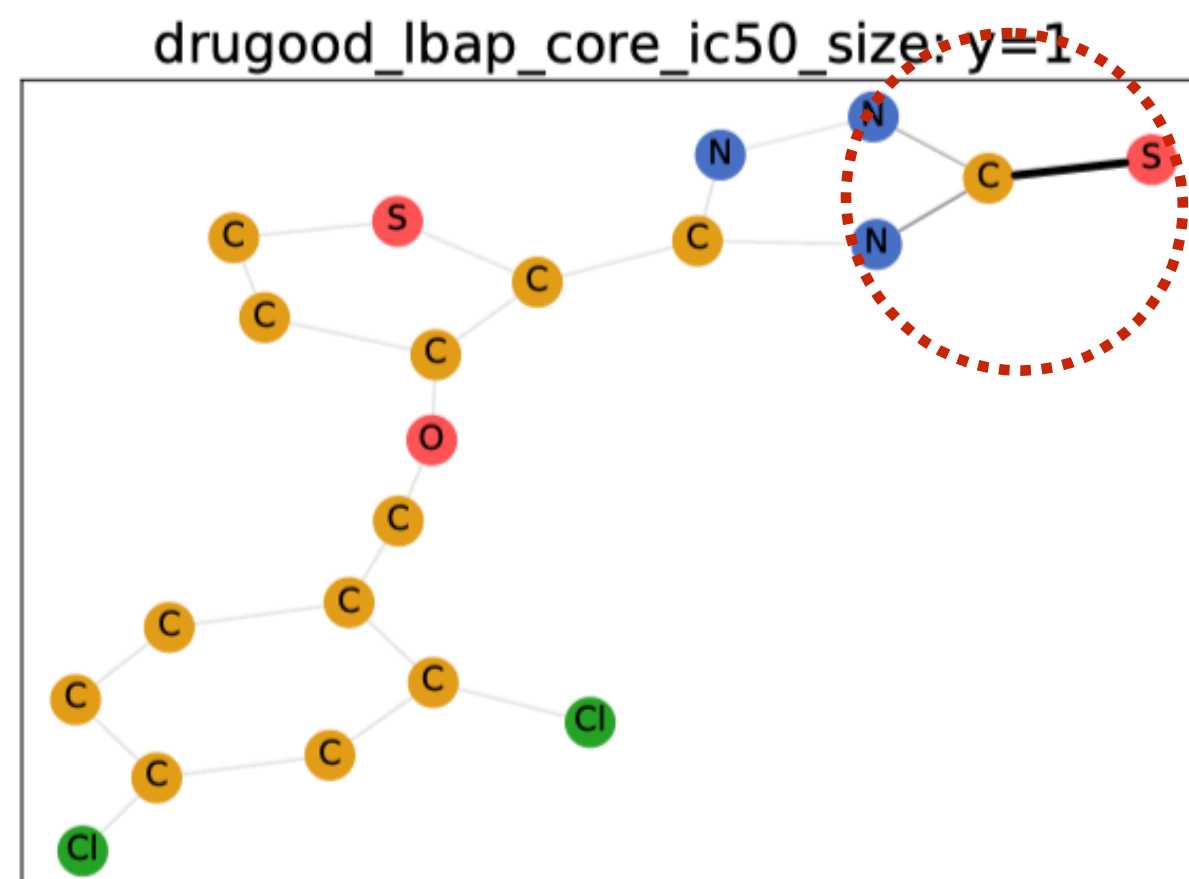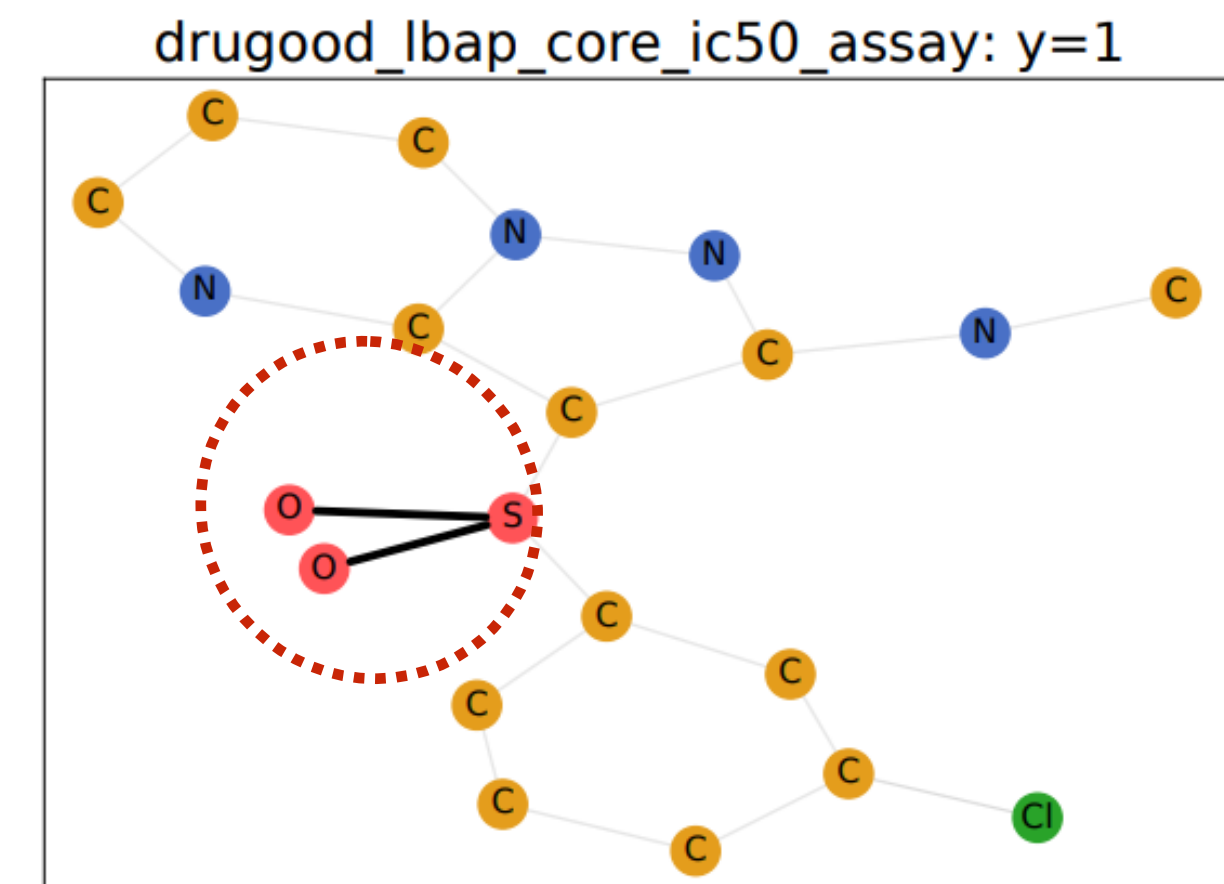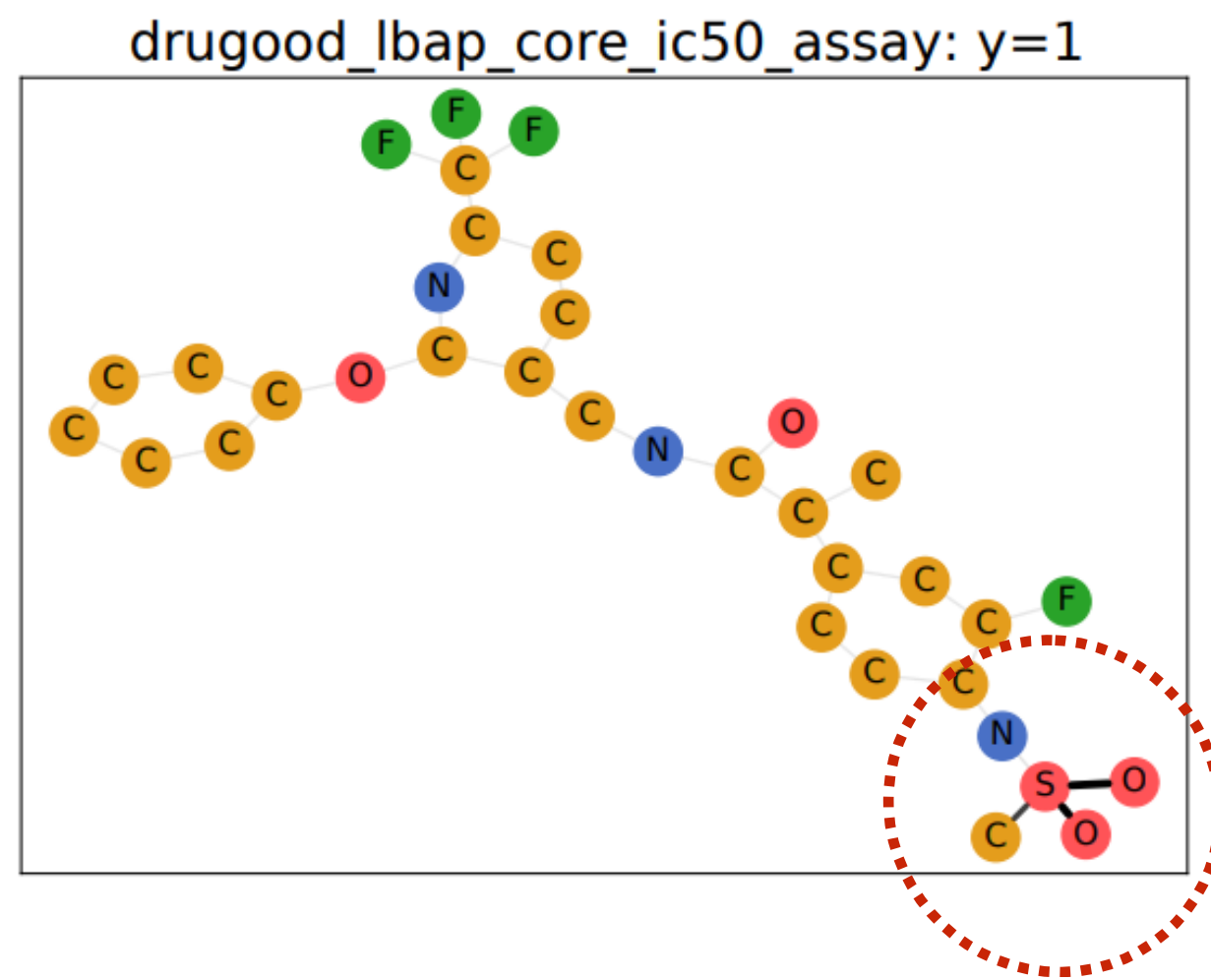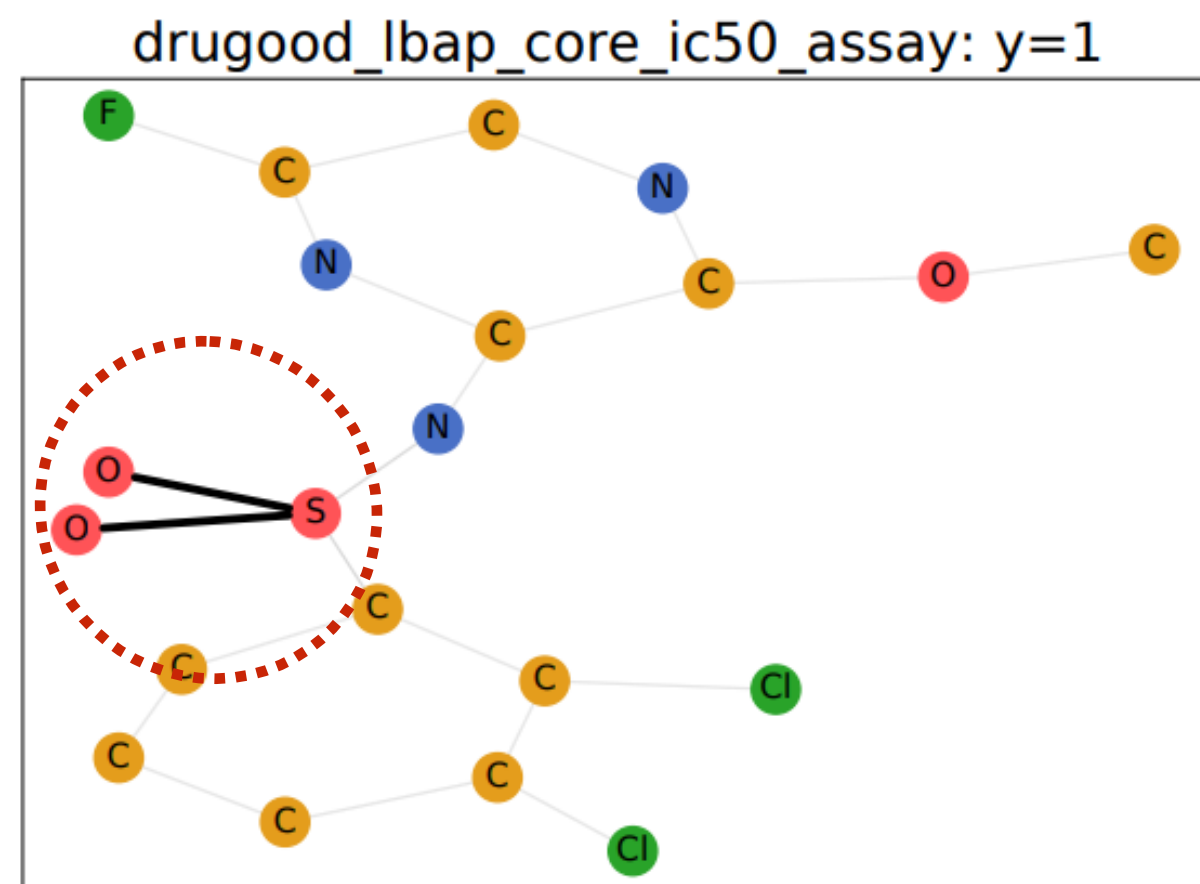
Table 3: OOD generalization performance on graph size shifts for real-world graphs in terms of Matthews correlation coefficient.

DATASETS	NCI1	NCI109	PROTEINS	DD	AVG
ERM	0.15 (0.05)	0.16 (0.02)	0.22 (0.09)	0.27 (0.09)	0.20
ASAP	0.16 (0.10)	0.15 (0.07)	0.22 (0.16)	0.21 (0.08)	0.19
GIB	0.13 (0.10)	0.16 (0.02)	0.19 (0.08)	0.01 (0.18)	0.12
DIR	0.21 (0.06)	0.13 (0.05)	0.25 (0.14)	0.20 (0.10)	0.20
IRM	0.17 (0.02)	0.14 (0.01)	0.21 (0.09)	0.22 (0.08)	0.19
V-REX	0.15 (0.04)	0.15 (0.04)	0.22 (0.06)	0.21 (0.07)	0.18
EIIL	0.14 (0.03)	0.16 (0.02)	0.20 (0.05)	0.23 (0.10)	0.19
IB-IRM	0.12 (0.04)	0.15 (0.06)	0.21 (0.06)	0.15 (0.13)	0.16
CNC	0.16 (0.04)	0.16 (0.04)	0.19 (0.08)	0.27 (0.13)	0.20
WL KERNEL	**0.39 (0.00)**	0.21 (0.00)	0.00 (0.00)	0.00 (0.00)	0.15
GC KERNEL	0.02 (0.00)	0.00 (0.00)	0.29 (0.00)	0.00 (0.00)	0.08
$\Gamma_{1\text{-HOT}}$	0.17 (0.08)	**0.25 (0.06)**	0.12 (0.09)	0.23 (0.08)	0.19
Γ_{GIN}	0.24 (0.04)	0.18 (0.04)	0.29 (0.11)	**0.28 (0.06)**	0.25
Γ_{RPGIN}	0.26 (0.05)	0.20 (0.04)	0.25 (0.12)	0.20 (0.05)	0.23
CIGAv1	0.22 (0.07)	**0.23 (0.09)**	**0.40 (0.06)**	**0.29 (0.08)**	**0.29**
CIGAv2	**0.27 (0.07)**	0.22 (0.05)	**0.31 (0.12)**	0.26 (0.08)	**0.27**
ORACLE (IID)	0.32 (0.05)	0.37 (0.06)	0.39 (0.09)	0.33 (0.05)	

CIGA outperforms previous methods under other **realistic shifts** by a significant margin up to **10%**.

Interpretable Studies of CIGA

CIGA finds interesting critical functional groups/sub-molecules in OOD molecular affinity prediction.

(Ji et al., 2022;)

Summary

Through the lens of causality, we establish general SCMs to characterize the distribution shifts on graphs, and generalize the invariance principle to graphs.

We instantiate the invariance principle through a novel framework CIGA, where the prediction is decomposed into the subgraph identification and classification.

We show that the provable identification of the underlying invariant subgraph can be achieved using a contrastive strategy both theoretically and empirically.

Paper   Code

Thank you!

Contact: yqchen@cse.cuhk.edu.hk