



TEXAS A&M UNIVERSITY  
Engineering

# TOWARDS SELF-SUPERVISED LEARNING AND EXPLAINING OF DEEP MODELS

Yaochen Xie

**Advisor:** Dr. Shuiwang Ji

**Advisory Committee Members**

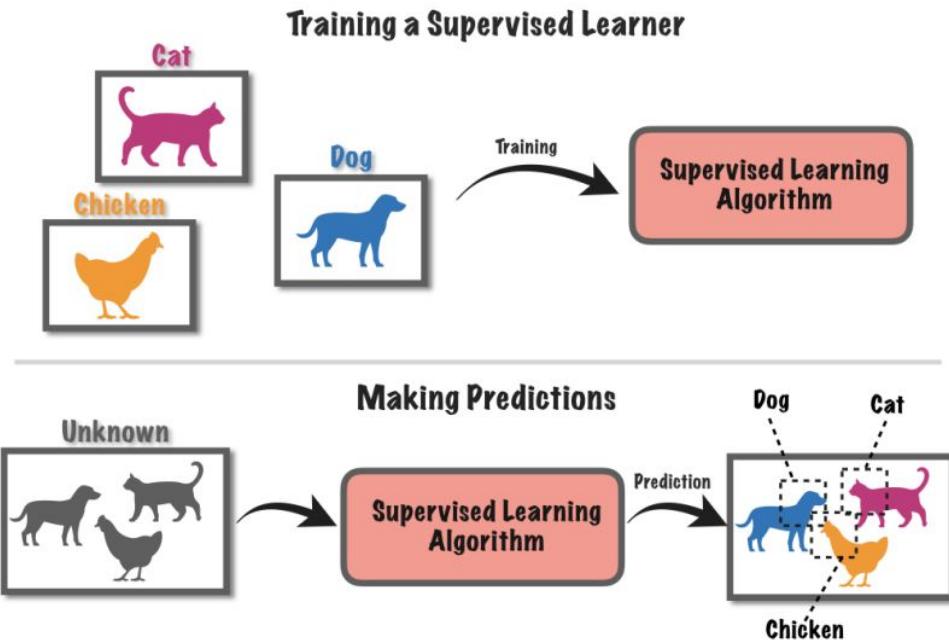
Dr. Yu Ding  
Dr. Ruihong Huang  
Dr. Bobak Mortazavi

# Supervised Learning



TEXAS A&M UNIVERSITY  
Engineering

- The most common learning setting for deep models
- Requires data with paired labels

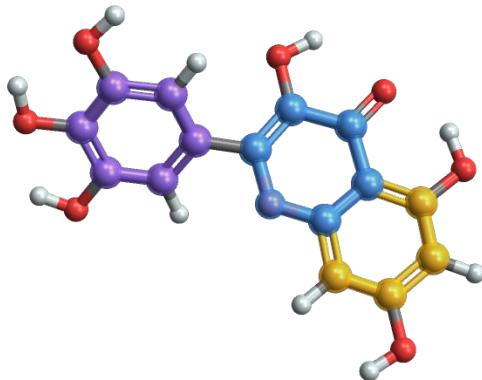


<https://towardsdatascience.com/supervised-vs-unsupervised-learning-in-2-minutes-72dad148f242>

# Labels Are Expensive/Unavailable When ...

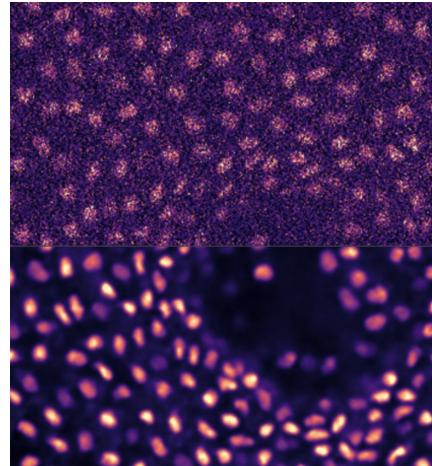


TEXAS A&M UNIVERSITY  
Engineering



## Drug Discovery

<https://www.wolfram.com/language/12/molecular-structure-and-computation/molecule-graphs.html.en?product=mathematica>



## Augmented Microscopy

Weigert, Martin, et al. "Content-aware image restoration: pushing the limits of fluorescence microscopy." *Nature methods* 15.12 (2018): 1090.



## Industrial Large-Scale Models

<https://www.fincash.com/l/basics/ga-fam-stocks>

How about if we...

## Drug Discovery

Learn from a huge amount  
of unlabeled molecules?

## Augmented Microscopy

Perform denoising without  
ground-truth images?

## Industrial Large-Scale Models

Learn an encoding model  
for general purpose?

How about if we...

## Drug Discovery

Learn from a huge amount  
of unlabeled molecules?

## Augmented Microscopy

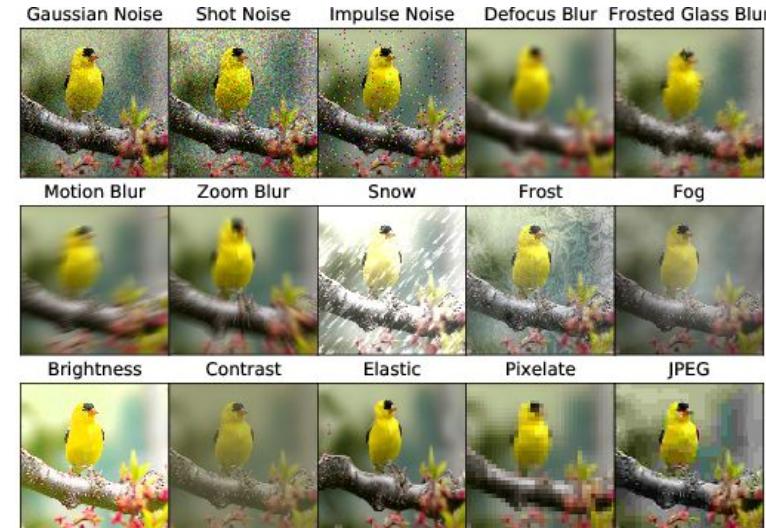
Perform denoising without  
ground-truth images?

## Industrial Large-Scale Models

Learn an encoding model  
for general purpose?

We can involve the **supervision** from data **itself!**

- A general goal of deep learning: to learn good representation to perform a given task
- What are good representations?
  - Informative
  - Robust to noise
- Supervised v.s. Self-supervised



Hendrycks and Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." ICLR 2019.

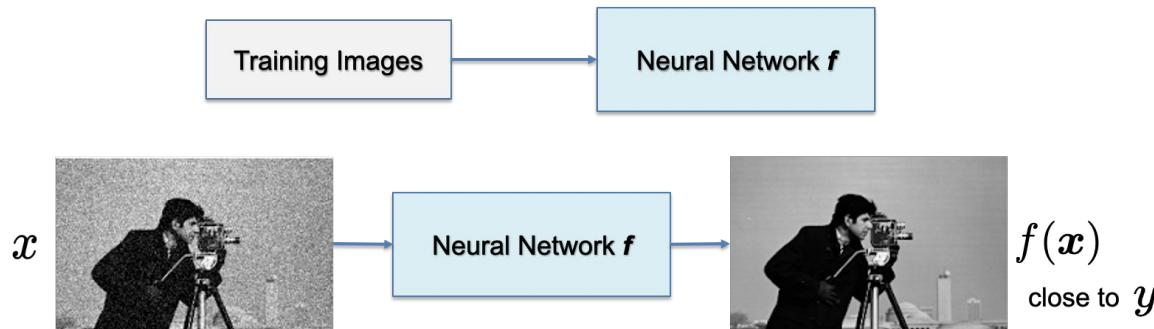


- Self-supervised image denoising
- Self-supervised graph representation learning
- Self-supervised explainer models
- Self-supervised learning for genetics studies



- Self-supervised image denoising
- Self-supervised graph representation learning
- Self-supervised explainer models
- Self-supervised learning for genetics studies

- The most common setting to training a learning-based denoising model
  - Samples of paired noisy-clean images ( $x_i, y_i$ ) are given as the training data.
  - Training with the supervised loss, e.g., MSE.



Xie et al., Noise2Same: Optimizing a self-supervised bound for image denoising. NeurIPS'20.

- The most common setting to training a learning-based denoising model
  - Samples of paired noisy-clean images ( $x_i, y_i$ ) are given as the training data.
  - Training with the supervised loss, e.g., MSE.

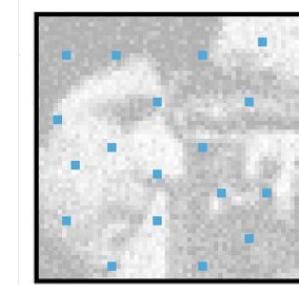
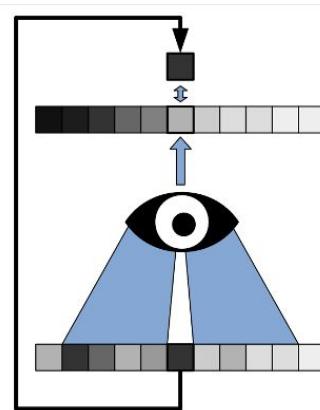
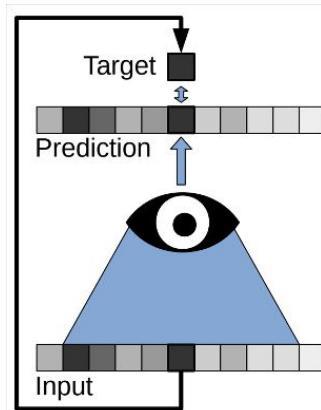
Clean images are usually unavailable in most real world scenarios!

Self-supervised denoising: an even more strict setting of available data

- Only (unpaired) noisy images  $x$  are given in the training dataset.
- How to design a task to utilize  $x$  as both the source and the target, with certain restrictions to avoid learning identical mapping.

Self-supervised denoising: an even more strict setting of available data

- Only (unpaired) noisy images  $x$  are given in the training dataset.
- **Blind-spot denoising**



Krull A et al. *Noise2void-learning denoising from single noisy images*. CVPR 2019.

## Blind-spot denoising

- Theoretical framework: a relationship between losses

$$\mathbb{E}_x \|f(\mathbf{x}) - \mathbf{x}\|^2 = \mathbb{E}_{x,y} \|f(\mathbf{x}) - \mathbf{y}\|^2 + \mathbb{E}_{x,y} \|\mathbf{x} - \mathbf{y}\|^2 - 2 \langle f(\mathbf{x}) - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle$$

*Self-supervised loss  
for blind-spot-based  
methods*

*Supervised loss*

*Is zero when assuming  
f is “J-invariant”*

- Basic idea: using a self-supervised loss to approximate the supervised denoising.

**Xie et al., Noise2Same: Optimizing a self-supervised bound for image denoising.** NeurIPS'20.

## Blind-spot denoising

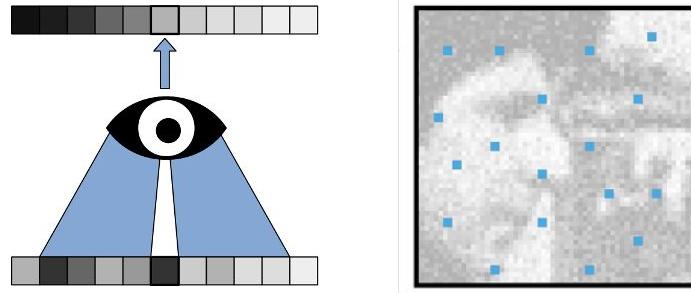
- Theoretical framework: a relationship between losses

$$\mathbb{E}_x \|f(x) - x\|^2 = \mathbb{E}_{x,y} \|f(x) - y\|^2 + \mathbb{E}_{x,y} \|x - y\|^2 - 2 \langle f(x) - y, x - y \rangle$$

*Self-supervised loss  
for blind-spot-based  
methods*

*Supervised loss*

*Is zero when assuming  
f is “J-invariant”*



- Issues?

Xie et al., Noise2Same: Optimizing a self-supervised bound for image denoising. NeurIPS'20.

- Trains the denoising model without assuming “J -invariant”  $f$ .
- Utilizes information of center pixel without knowing the noise model or post-processing. And hence can be applied in broader scenarios.

$$\mathbb{E}_x \|f(x) - x\|^2 = \mathbb{E}_{x,y} \|f(x) - y\|^2 + \mathbb{E}_{x,y} \|x - y\|^2 - 2 \langle f(x) - y, x - y \rangle$$

**Part of the**  
*Self-supervised loss*  
*in Noise2Same*

*Supervised loss*

We **bound this term** in  
a self-supervised  
fashion, instead of  
assuming “J -invariant”  
and making it zero

- We derive a **self-supervised upper bound** of the MSE loss

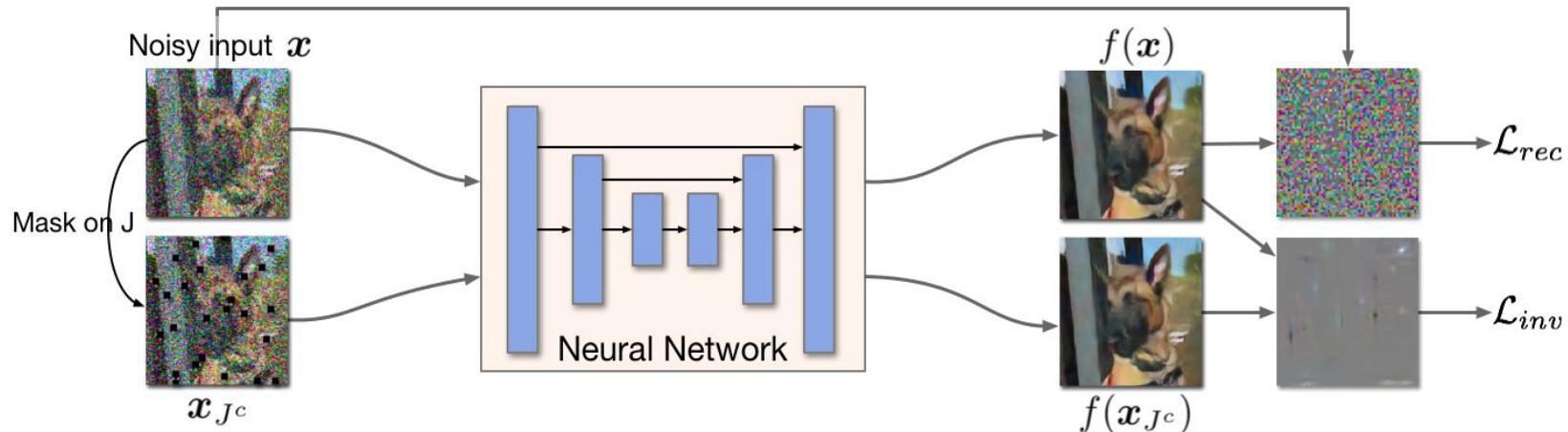
**Theorem 1.** Consider a normalized noisy image  $\mathbf{x} \in \mathbb{R}^m$  (obtained by subtracting the mean and dividing by the standard deviation) and its ground truth signal  $\mathbf{y} \in \mathbb{R}^m$ . Assume the noise is zero-mean and i.i.d among all the dimensions, and let  $J$  be a subset of  $m$  dimensions uniformly sampled from the image  $\mathbf{x}$ . For any  $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$ , we have

$$\mathbb{E}_{x,y} \|f(\mathbf{x}) - \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 \leq \mathbb{E}_x \|f(\mathbf{x}) - \mathbf{x}\|^2 + 2m \mathbb{E}_J \left[ \frac{\mathbb{E}_x \|f(\mathbf{x})_J - f(\mathbf{x}_{J^c})_J\|^2}{|J|} \right]^{1/2} \quad (6)$$

Supervised loss

**The self-supervised bound as the new loss**

- It is simple to implement the self-supervised bound  $L = L_{rec} + \lambda_{inv}\sqrt{L_{inv}}$



Xie et al., Noise2Same: Optimizing a self-supervised bound for image denoising. NeurIPS'20.

## How does the invariance term work?

- We derive a tighter upper bound assuming additive noise with known variance.
- The invariance term controls **how much information is utilized from center pixels**. When the noise is stronger, less center pixel information is used.

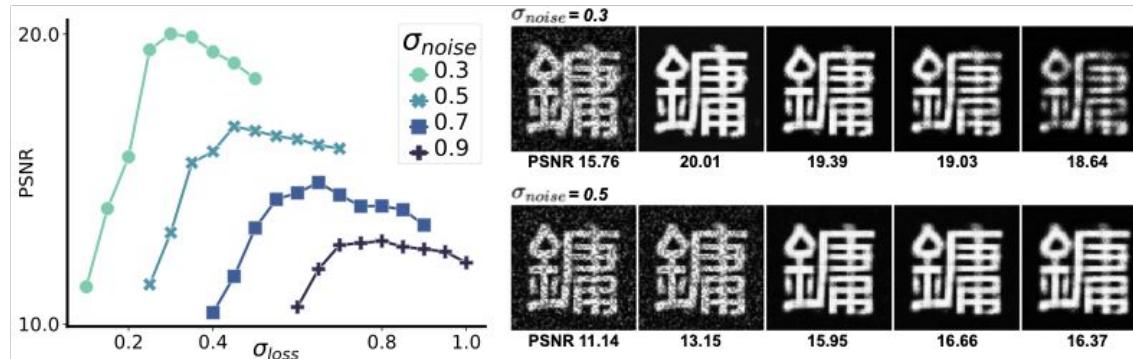
**Theorem 2.** *Given a noisy image  $\mathbf{x} \in \mathbb{R}^m$  and its ground truth signal  $\mathbf{y}$ , assuming the noise is independent among all dimensions, and further assuming the noise is additive with standard deviation  $\sigma$  and zero-mean, letting  $J$  be a subset of  $m$  dimensions uniformly sampled from the image  $\mathbf{x}$ , we have*

$$\mathbb{E}_{x,y} \|f(\mathbf{x}) - \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 \leq \mathbb{E}_x \|f(\mathbf{x}) - \mathbf{x}\|^2 + 2m\sigma \mathbb{E}_J \left[ \mathbb{E} \|f(\mathbf{x})_J - f(\mathbf{x}_{J^c})_J\|^2 / |J| \right]^{1/2}$$

Xie et al., Noise2Same: Optimizing a self-supervised bound for image denoising. NeurIPS'20.

## How does the invariance term work?

- We derive a tighter upper bound assuming additive noise with known variance.
- The invariance term controls **how much information is utilized from center pixels**.  
When the noise is stronger, less center pixel information is used.



Xie et al., Noise2Same: Optimizing a self-supervised bound for image denoising. NeurIPS'20.

# Experimental Results



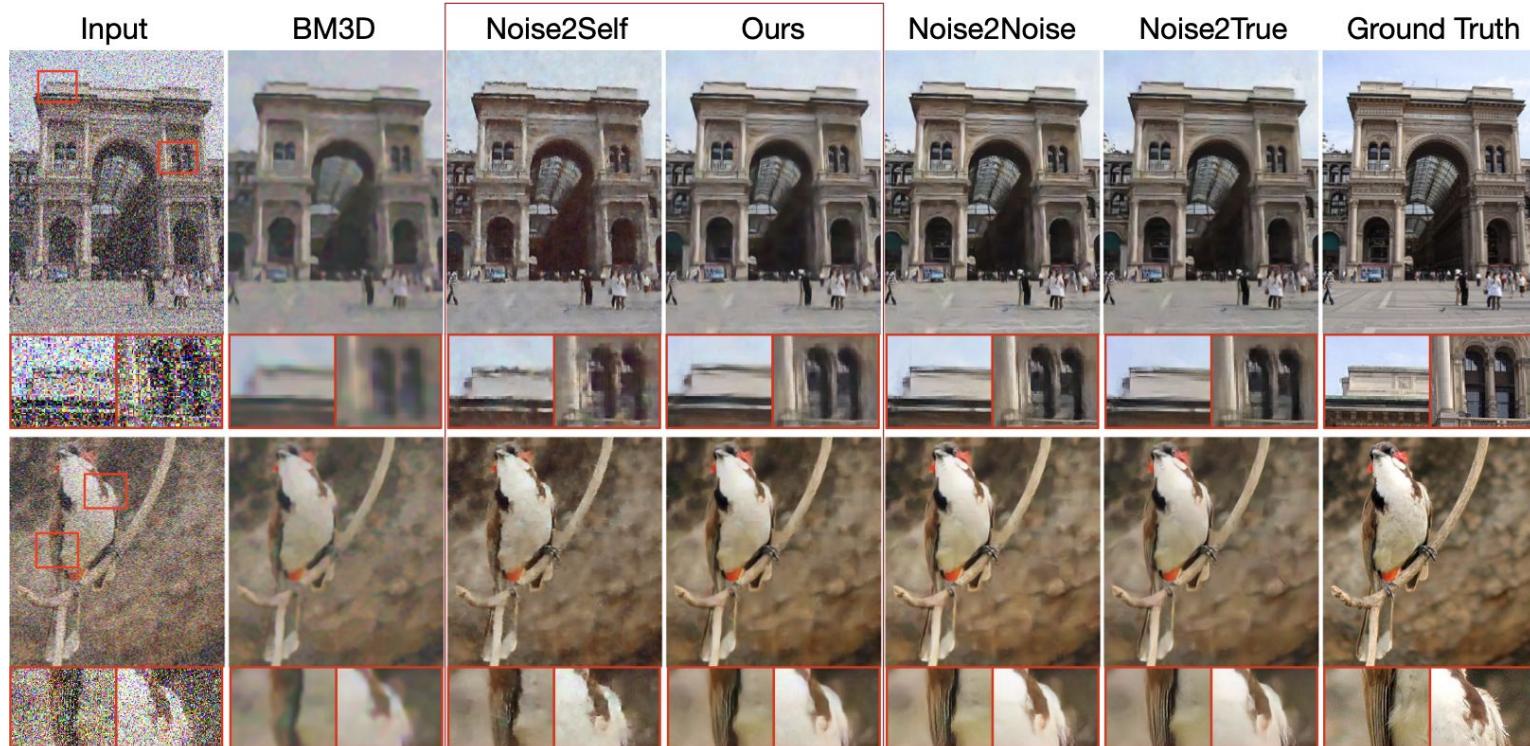
|                                      | <b>Methods</b>       | <b>Datasets</b> |              |                              |              |
|--------------------------------------|----------------------|-----------------|--------------|------------------------------|--------------|
|                                      |                      | ImageNet        | HànZì        | Planaria                     | BSD68        |
| <i>Traditional</i>                   | Input                | 9.69            | 6.45         | 21.52 / 21.09 / 20.82        | 20.19        |
|                                      | NLM [3]              | 18.04           | 8.41         | 25.80 / 24.03 / 21.62        | 22.73        |
|                                      | BM3D [5]             | 18.74           | 10.90        | -                            | 28.59        |
| <i>Supervised</i>                    | Noise2True           | 23.39           | 15.66        | 31.57 / 30.15 / 28.13        | 29.06        |
|                                      | Noise2Noise [13]     | 23.27           | 14.30        | -                            | 28.86        |
| <i>Self-Supervised + noise model</i> | Laine et al. [12]    | -               | -            | -                            | 28.84        |
| <i>Self-Supervised</i>               | Laine et al. [12]    | 20.89           | 10.70        | -                            | 27.15        |
|                                      | Noise2Void [10]      | 21.36           | 13.72        | 25.84 / 23.57 / 21.60        | 27.71        |
|                                      | Noise2Self-Noise [1] | 20.38           | 13.94        | 27.58 / 24.83 / 21.83        | 26.98        |
|                                      | Noise2Self-Donut [1] | 8.62            | 13.29        | 27.63 / 24.72 / 21.73        | <b>28.20</b> |
|                                      | <b>Noise2Same</b>    | <b>22.26</b>    | <b>14.38</b> | <b>29.48 / 26.93 / 22.41</b> | 27.95        |

Xie et al., Noise2Same: Optimizing a self-supervised bound for image denoising. NeurIPS'20.

# Experimental Results



TEXAS A&M UNIVERSITY  
Engineering



Xie et al., Noise2Same: Optimizing a self-supervised bound for image denoising. NeurIPS'20.



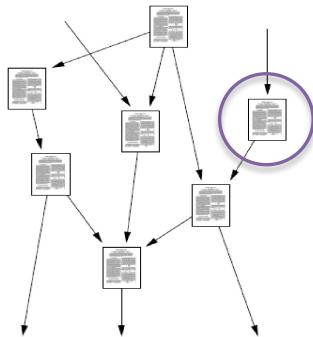
- Self-supervised image denoising
- Self-supervised graph representation learning
- Self-supervised explainer models
- Self-supervised learning for genetics studies

# Self-supervised graph representation learning

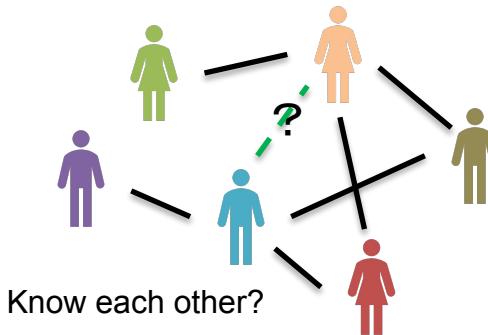
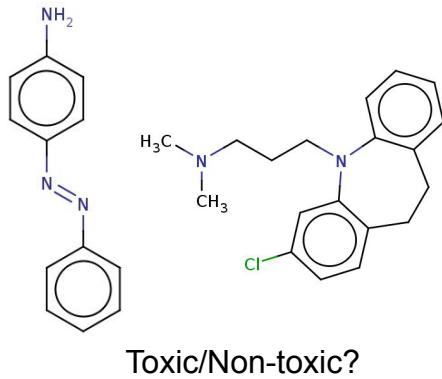


TEXAS A&M UNIVERSITY  
Engineering

- Graph data



Computer  
science?  
Math?  
Chemistry?

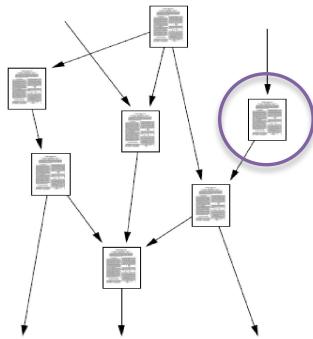


# Self-supervised graph representation learning

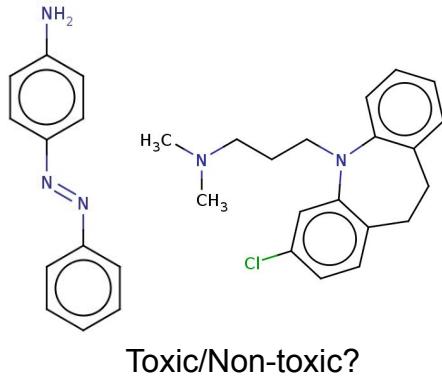


TEXAS A&M UNIVERSITY  
Engineering

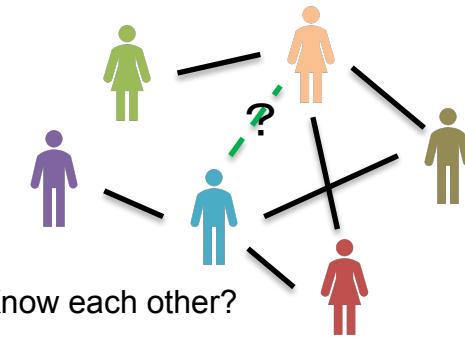
- Graph data



Computer  
science?  
Math?  
Chemistry?



Toxic/Non-toxic?



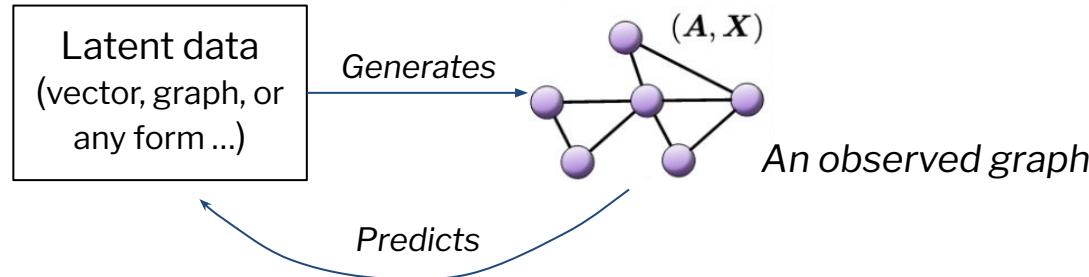
Know each other?

**Label acquisition/annotation is even more challenging for graph data.**

# Key Idea: Latent Data



- We consider the concept latent data, where any observed graph  $G = (A, X)$  is generated from a corresponding latent data that determine its semantic.



- We consider the concept latent data, where any observed graph  $G = (A, X)$  is generated from a corresponding latent data that determine its semantic.
- WLOG, we specifically consider latent data  $G_\ell = (A, F)$  in graph-structure with the same connectivity and satisfying two assumptions (non-structural and unbiased noise).
- Theorems can be generalized with other distances when considering latent data in different forms.

- We adopt the prediction/reconstruction of the latent graph to derive our predictive SSL task.

$$f^* = \arg \min_f \mathbb{E} \|f(\mathbf{A}, \mathbf{X}) - \mathbf{F}\|^2$$

- We derive a self-supervised upper bound for the above objective to eliminate the need of unknown  $\mathbf{F}$

$$\begin{aligned} \mathbb{E}_{\mathbf{A}, \mathbf{X}, \mathbf{F}} \left[ \|f(\mathbf{A}, \mathbf{X}) - \mathbf{F}\|^2 + \|\mathbf{X} - \mathbf{F}\|^2 \right] &\leq \mathbb{E}_{\mathbf{A}, \mathbf{X}} \|f(\mathbf{A}, \mathbf{X}) - \mathbf{X}\|^2 + \\ &2\sigma|V| \mathbb{E}_J \left[ \frac{\mathbb{E}_{\mathbf{A}, \mathbf{X}} \|f_J(\mathbf{A}, \mathbf{X}) - f_J(\mathbf{A}, \mathbf{X}_{J^c})\|^2}{|J|} \right]^{1/2} \end{aligned}$$

## Node-level representation learning

**Corollary 2.2.** *Let  $G = (\mathbf{A}, \mathbf{X})$  be a given graph,  $G_{\mathcal{I}} = (\mathbf{A}, \mathbf{F})$  be its latent graph,  $\mathcal{E}$  and  $\mathcal{D}$  be a graph encoder and a prediction head (decoder) consisting of fully-connected layers. If the prediction head  $\mathcal{D}$  is  $\ell$ -Lipschitz continuous with respect to  $l_2$ -norm, we further have the following inequality,*

$$\begin{aligned} \mathbb{E} [\|\mathcal{D}(\mathbf{H}) - \mathbf{F}\|^2 + \|\mathbf{X} - \mathbf{F}\|^2] &\leq \mathbb{E} \|\mathcal{D}(\mathbf{H}) - \mathbf{X}\|^2 \\ &+ 2\sigma|V|\ell \mathbb{E}_J \left[ \frac{\mathbb{E} \|\mathbf{H}_J - \mathbf{H}'_J\|^2}{|J|} \right]^{1/2}, \end{aligned} \quad (3)$$

where  $\mathbf{H} = \mathcal{E}(\mathbf{A}, \mathbf{X})$  and  $\mathbf{H}' = \mathcal{E}(\mathbf{A}, \mathbf{X}_{J^c})$  denote the node embedding of the given graph and the masked graph, respectively, and  $\mathbf{H}_J := \mathbf{H}[J, :]$  selects rows with indices in  $J$ .

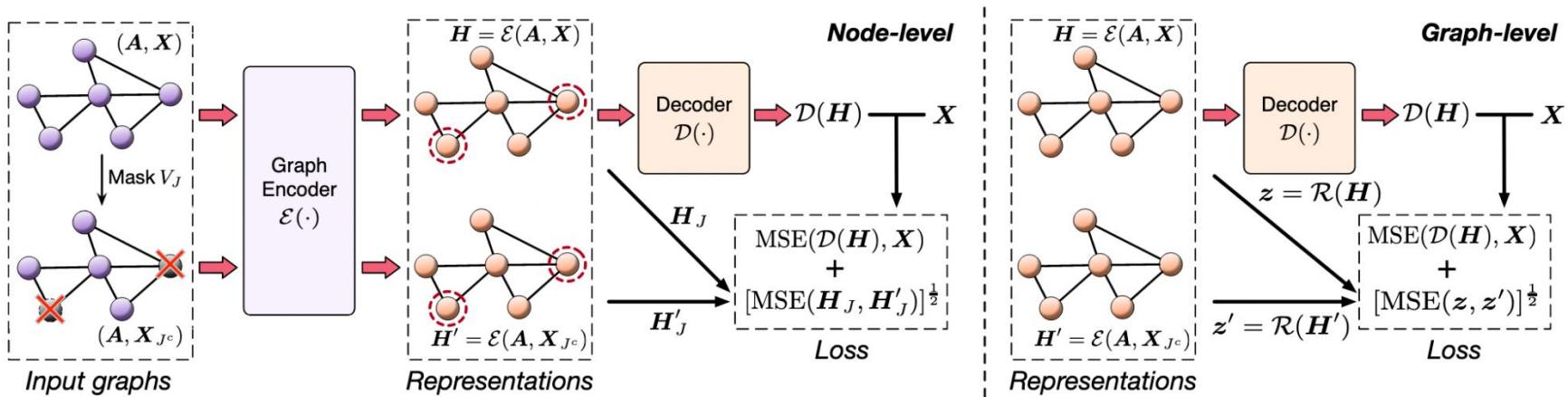
## Graph-level representation learning

**Corollary 2.3.** *Let  $G = (\mathbf{A}, \mathbf{X})$  be a given graph,  $G_{\mathcal{I}} = (\mathbf{A}, \mathbf{F})$  be its hidden latent graph,  $\mathcal{E}$  be a graph encoder,  $\mathcal{R}$  be a readout function satisfying  $k$ -Bilipschitz continuity with respect to  $l_2$ -norm, and  $\mathcal{D}$  be a prediction head (decoder). If the prediction head  $\mathcal{D}$  is  $\ell$ -Lipschitz continuous with respect to  $l_2$ -norm, we have the following inequality,*

$$\begin{aligned} \mathbb{E} [\|\mathcal{D}(\mathbf{H}) - \mathbf{F}\|^2 + \|\mathbf{X} - \mathbf{F}\|^2] &\leq \mathbb{E} \|\mathcal{D}(\mathbf{H}) - \mathbf{X}\|^2 \\ &+ 2\sigma|V|k\ell \mathbb{E}_J \left[ \frac{\mathbb{E} \|\mathbf{z} - \mathbf{z}'\|^2}{|J|} \right]^{1/2}, \end{aligned} \quad (4)$$

where  $\mathbf{z} = \mathcal{R}(\mathbf{H})$  and  $\mathbf{z}' = \mathcal{R}(\mathbf{H}')$  denote the graph-level representations of the given graph and the masked graph, respectively.

# The LaGraph Framework



Xie et al., Self-Supervised Representation Learning via Latent Graph Prediction. ICML'22.

# The Information Bottleneck Principle



TEXAS A&M UNIVERSITY  
Engineering

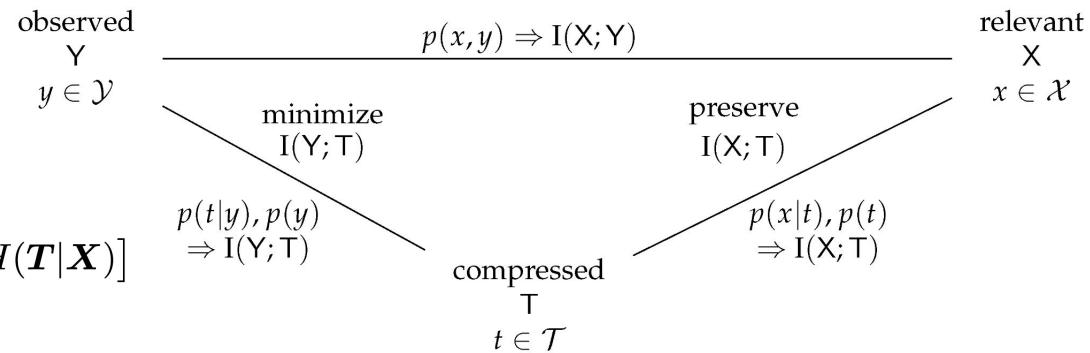
Intuition: a good representation (compression) of data should be

- I. As informative as possible
- II. Robust to noise in the observed data

$$\mathbf{T}^* = \arg \min_{\mathbf{T}} I(\mathbf{T}; \mathbf{Y}) - \beta I(\mathbf{T}; \mathbf{X})$$



$$\begin{aligned}\mathbf{T}^* &= \arg \min_{\mathbf{T}} [H(\mathbf{T}) - H(\mathbf{T}|\mathbf{Y})] \\ &\quad - \beta [H(\mathbf{T}) - H(\mathbf{T}|\mathbf{X})] \\ &= \arg \min_{\mathbf{T}} H(\mathbf{T}|\mathbf{X}) - \lambda H(\mathbf{T}),\end{aligned}$$



Xie et al., Self-Supervised Representation Learning via Latent Graph Prediction. ICML'22.

# The Information Bottleneck Principle



**Reconstruction term:** Computed between original node features  $X$ , to ensure the learning of **informative** representations.

**Invariance term:** Computed between representations of original graph and masked graph respectively, to ensure the learning of **non-trivial** representations (e.g., to avoid identical mapping).

**Corollary 2.2.** Let  $G = (\mathbf{A}, \mathbf{X})$  be a given graph,  $G_{\mathcal{I}} = (\mathbf{A}, \mathbf{F})$  be its latent graph,  $\mathcal{E}$  and  $\mathcal{D}$  be a graph encoder and a prediction head (decoder) consisting of fully-connected layers. If the prediction head  $\mathcal{D}$  is  $\ell$ -Lipschitz continuous with respect to  $l_2$ -norm, we further have the following inequality,

$$\mathbb{E} [\|\mathcal{D}(\mathbf{H}) - \mathbf{F}\|^2 + \|\mathbf{X} - \mathbf{F}\|^2] \leq \mathbb{E} \|\mathcal{D}(\mathbf{H}) - \mathbf{X}\|^2 + 2\sigma|V|\ell \mathbb{E}_J \left[ \frac{\mathbb{E} \|\mathbf{H}_J - \mathbf{H}'_J\|^2}{|J|} \right]^{1/2}, \quad (3)$$

where  $\mathbf{H} = \mathcal{E}(\mathbf{A}, \mathbf{X})$  and  $\mathbf{H}' = \mathcal{E}(\mathbf{A}, \mathbf{X}_{J^c})$  denote the node embedding of the given graph and the masked graph, respectively, and  $\mathbf{H}_J := \mathbf{H}[J, :]$  selects rows with indices in  $J$ .

# Results: Node-level Tasks



*Table 2.* Performance on node-level datasets, 20 runs averaged. Results of SSL methods with the best performance are highlighted in bold numbers. *Left:* Mean classification accuracy on transductive datasets, with baseline results from Thakoor et al. (2021). *Right:* Micro-averaged F1 scores on larger-scale inductive datasets, with baseline results from Thakoor et al. (2021) and Jiao et al. (2020).

| Transductive | Am.Comp.                       | Am.Pht.                        | Co.CS                          | Co.Phy                         | Inductive | PPI                            | Flickr                         | Reddit                         |
|--------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|-----------|--------------------------------|--------------------------------|--------------------------------|
| Raw features | 73.8 $\pm$ 0.0                 | 78.5 $\pm$ 0.0                 | 90.4 $\pm$ 0.0                 | 93.6 $\pm$ 0.0                 | Raw feat. | 42.5 $\pm$ 0.3                 | 20.3 $\pm$ 0.2                 | 58.5 $\pm$ 0.1                 |
| DeepWalk     | 85.7 $\pm$ 0.1                 | 89.4 $\pm$ 0.1                 | 84.6 $\pm$ 0.2                 | 91.8 $\pm$ 0.2                 | GAE       | 75.7 $\pm$ 0.0                 | 50.7 $\pm$ 0.2                 | OOM                            |
| GAE          | 87.7 $\pm$ 0.3                 | 92.7 $\pm$ 0.3                 | 92.4 $\pm$ 0.2                 | 95.3 $\pm$ 0.1                 | VGAE      | 75.8 $\pm$ 0.0                 | 50.4 $\pm$ 0.2                 | OOM                            |
| VGAE         | 88.1 $\pm$ 0.3                 | 92.8 $\pm$ 0.3                 | 92.5 $\pm$ 0.2                 | 95.3 $\pm$ 0.1                 | Super-GCN | 51.5 $\pm$ 0.6                 | 48.7 $\pm$ 0.3                 | 93.3 $\pm$ 0.1                 |
| Supervised   | 86.5 $\pm$ 0.5                 | 92.4 $\pm$ 0.2                 | 93.0 $\pm$ 0.3                 | 95.7 $\pm$ 0.2                 | Super-GAT | 97.3 $\pm$ 0.2                 | OOM                            | OOM                            |
| DGI          | 84.0 $\pm$ 0.5                 | 91.6 $\pm$ 0.2                 | 92.2 $\pm$ 0.6                 | 94.5 $\pm$ 0.5                 | GraphSAGE | 46.5 $\pm$ 0.7                 | 36.5 $\pm$ 1.0                 | 90.8 $\pm$ 1.1                 |
| GMI          | 82.2 $\pm$ 0.3                 | 90.7 $\pm$ 0.2                 | OOM                            | OOM                            | DGI       | 63.8 $\pm$ 0.2                 | 42.9 $\pm$ 0.1                 | 94.0 $\pm$ 0.1                 |
| MVGRL        | 87.5 $\pm$ 0.1                 | 91.7 $\pm$ 0.1                 | 92.1 $\pm$ 0.1                 | 95.3 $\pm$ 0.0                 | GMI       | 65.0 $\pm$ 0.0                 | 44.5 $\pm$ 0.2                 | 95.0 $\pm$ 0.0                 |
| GRACE        | 87.5 $\pm$ 0.2                 | 92.2 $\pm$ 0.2                 | 92.9 $\pm$ 0.0                 | 95.3 $\pm$ 0.0                 | SUBG-CON  | 66.9 $\pm$ 0.2                 | 48.8 $\pm$ 0.1                 | <b>95.2<math>\pm</math>0.0</b> |
| GCA          | 88.9 $\pm$ 0.2                 | 92.5 $\pm$ 0.2                 | 93.1 $\pm$ 0.0                 | 95.7 $\pm$ 0.0                 | BGRL-GCN  | 69.6 $\pm$ 0.2                 | 50.0 $\pm$ 0.3*                | OOM*                           |
| BGRL         | <b>89.7<math>\pm</math>0.3</b> | 92.9 $\pm$ 0.3                 | 93.2 $\pm$ 0.2                 | 95.6 $\pm$ 0.1                 | BGRL-GAT  | 70.5 $\pm$ 0.1                 | 44.2 $\pm$ 0.1*                | OOM*                           |
| LaGraph      | 88.0 $\pm$ 0.3                 | <b>93.5<math>\pm</math>0.4</b> | <b>93.3<math>\pm</math>0.2</b> | <b>95.8<math>\pm</math>0.1</b> | LaGraph   | <b>74.6<math>\pm</math>0.0</b> | <b>51.3<math>\pm</math>0.1</b> | <b>95.2<math>\pm</math>0.0</b> |

# Results: Graph-level Tasks



TEXAS A&M UNIVERSITY  
Engineering

Table 1. Performance on graph-level classification tasks, scores are averaged over 5 runs. Bold and underlined numbers highlight the top-2 performance. OOM indicates running out-of-memory on a 56GB Nvidia A6000 GPU.

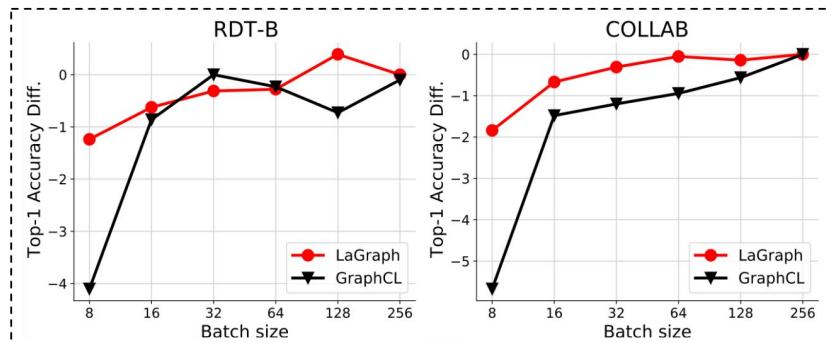
|           | NCI1            | PROTEINS        | DD              | MUTAG           | COLLAB          | RDT-B           | RDT-M5K         | IMDB-B          |
|-----------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| GL        | –               | –               | –               | 81.7±2.1        | –               | 77.3±0.2        | 41.0±0.2        | 65.9±1.0        |
| WL        | 80.0±0.5        | 72.9±0.6        | –               | 80.7±3.0        | –               | 68.8±0.4        | 46.1±0.2        | 72.3±3.4        |
| DGK       | 80.3±0.5        | 73.3±0.8        | –               | 87.4±2.7        | –               | 78.0±0.4        | 41.3±0.2        | 67.0±0.6        |
| Node2Vec  | 54.9±1.6        | 57.5±3.6        | 75.1±0.5        | 72.6±10.2       | 55.7±0.2        | 73.8±0.5        | 34.1±0.4        | 50.0±0.8        |
| Sub2Vec   | 52.8±1.5        | 53.0±5.6        | 73.6±1.5        | 61.1±15.8       | 62.1±1.4        | 71.5±0.4        | 36.7±0.4        | 55.3±1.5        |
| Graph2Vec | 73.2±1.8        | 73.3±2.1        | 76.2±0.1        | 83.2±9.3        | 59.9±0.0        | 75.8±1.0        | 47.9±0.3        | 71.1±0.5        |
| GAE       | 73.3±0.6        | 74.1±0.5        | 77.9±0.5        | 84.0±0.6        | 56.3±0.1        | 74.8±0.2        | 37.6±1.6        | 52.1±0.2        |
| VGAE      | 73.7±0.3        | 74.0±0.5        | 77.6±0.4        | 84.4±0.6        | 56.3±0.0        | 74.8±0.2        | 39.1±1.6        | 52.1±0.2        |
| InfoGraph | 76.2±1.1        | <u>74.4±0.3</u> | 72.9±1.8        | 89.0±1.1        | 70.7±1.1        | 82.5±1.4        | 53.5±1.0        | 73.0±0.9        |
| GraphCL   | <u>77.9±0.4</u> | 74.4±0.5        | <b>78.6±0.4</b> | 86.8±1.3        | 71.4±1.2        | <u>89.5±0.8</u> | <u>56.0±0.3</u> | 71.1±0.4        |
| MVGRL     | 75.1±0.5        | 71.5±0.3        | OOM             | <u>89.7±1.1</u> | OOM             | 84.5±0.6        | OOM             | <b>74.2±0.7</b> |
| LaGraph   | <b>79.9±0.5</b> | <b>75.2±0.4</b> | 78.1±0.4        | <b>90.2±1.1</b> | <b>77.6±0.2</b> | <b>90.4±0.8</b> | <b>56.4±0.4</b> | 73.7±0.9        |

# Scalability and Robustness

|        | # nodes sampled           | 100    | 1,000  | 2,500  | 5,000  | 10,000  | all     |
|--------|---------------------------|--------|--------|--------|--------|---------|---------|
| Flickr | % nodes sampled           | 0.22%  | 2.24%  | 5.60%  | 11.20% | 22.41%  | 100.00% |
|        | F1-score - <i>LaGraph</i> | 6.07   | 51.12  | 51.12  | 51.27  | 51.29   | 51.26   |
|        | Memory - <i>LaGraph</i>   | 1389MB | 1465MB | 1553MB | 1725MB | 2065MB  | 4211MB  |
|        | F1-score - GraphCL        | 45.27  | 45.27  | 45.27  | 45.38  | 45.45   | 45.48   |
|        | Memory - GraphCL          | 1647MB | 2599MB | 4137MB | 6741MB | 11905MB | 47939MB |
| Reddit | % nodes sampled           | 0.07%  | 0.65%  | 1.63%  | 3.25%  | 6.50%   | 100.00% |
|        | F1-score - <i>LaGraph</i> | 5.76   | 95.05  | 95.06  | 95.08  | 95.09   | 95.22   |
|        | Memory - <i>LaGraph</i>   | 1403MB | 1475MB | 1585MB | 1783MB | 2161MB  | 16933MB |
|        | F1-score - GraphCL        | 93.24  | 93.24  | 93.25  | 93.31  | 93.32   | OOM     |
|        | Memory - GraphCL          | 4199MB | 6117MB | 6687MB | 9297MB | 14495MB | OOM     |

↑ *Model scalability on large-scale graphs: less memory usage during training.*

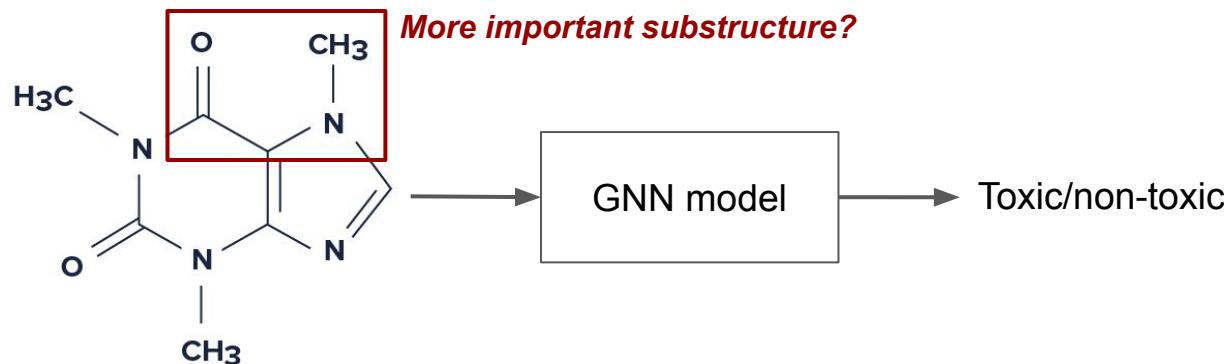
→ *Model robustness to small mini-batch sizes: less performance losses when memory is limited.*



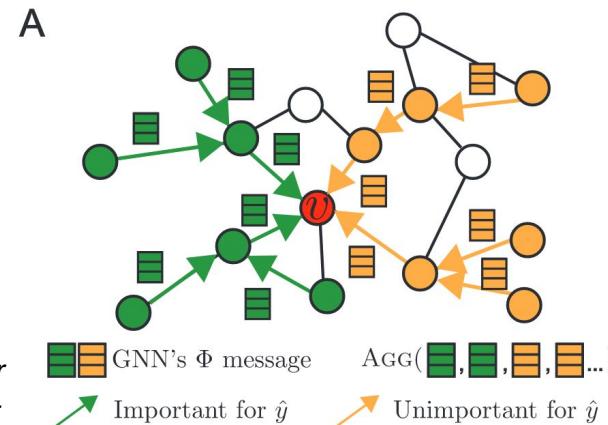


- Self-supervised image denoising
- Self-supervised graph representation learning
- Self-supervised explainer models
- Self-supervised learning for genetics studies

- Given a well-trained GNN and a graph instance, figure out the substructure that lead (contribute the most) to the prediction results.



- Perturbation-based explanation (e.g., GNNExplainer, PGExplainer).
- **Key idea:** to learn masks on given graph edges (and node attributes), s.t., when unimportant edges/attributes are masked-out, the prediction results of the GNN should be similar.
- Requires the entire GNN model to learn an explainer.



*Ying et al. GNNExplainer: Generating Explanations for Graph Neural Networks. NeurIPS 2019.*

- The 2-stage training is common in industrial applications.
- Issues: undefined downstream task and unavailable downstream models during stage 1, current explainers:
  - Unable to explain the GNN encoder **without downstream tasks**.
  - Unable to explain GNN encoder with **non-differentiable downstream models**.
  - Have to repeatedly train explainers **when downstream models are updated**.

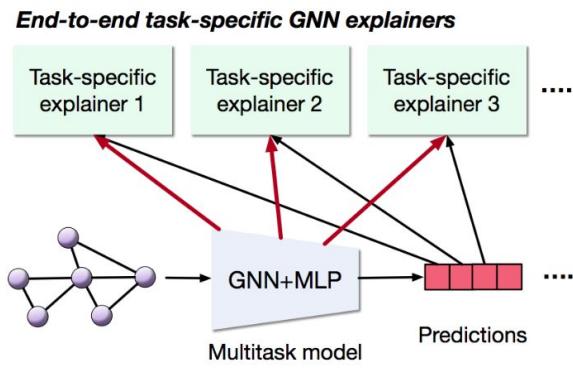
- Multi-task prediction is common for graph machine learning, e.g., to predict multiple chemical properties in drug discovery for a molecular graph.
  - ToxCast from MoleculeNet has 167 graph-level prediction tasks.
  - PPI has 121 node-level binary classification tasks.

|                                   | Learning | Inductive | Task-agnostic | # explainers required |
|-----------------------------------|----------|-----------|---------------|-----------------------|
| Gradient- & Rule-based            | No       | -         | -             | 1                     |
| GNNEExplainer (Ying et al., 2019) | Yes      | No        | No            | $M * N$               |
| SubgraphX (Yuan et al., 2021)     | Yes      | No        | No            | $M * N$               |
| PGEExplainer (Luo et al., 2020)   | Yes      | Yes       | No            | $M$                   |
| Task-agnostic explainers          | Yes      | Yes       | Yes           | 1                     |

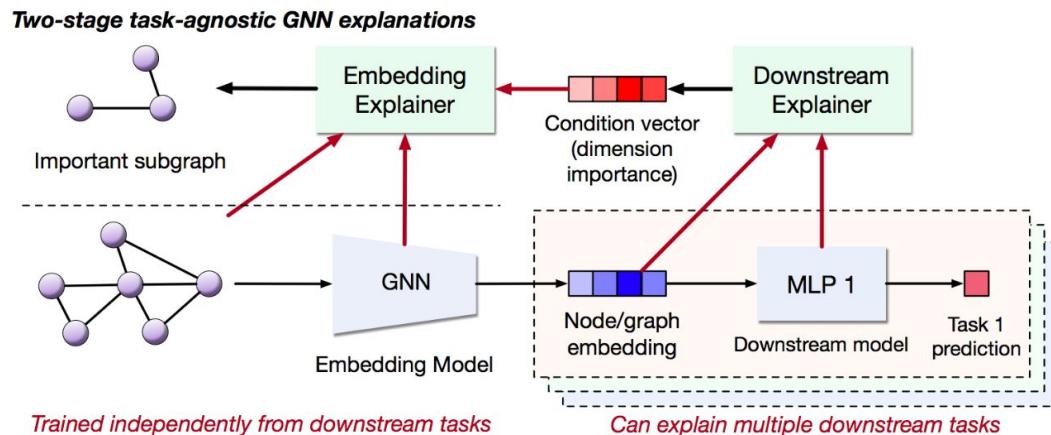
# The Task-Agnostic Graph Explanations



- To construct embedding explainer and downstream explainer respectively.



Need train different explainers to explain a multitask prediction model. Unable to train without downstream.



Trained independently from downstream tasks

Can explain multiple downstream tasks

- A simple gradient-based explainer, since the downstream models are usually based on fully-connected layers and are less challenging to explain.

$$\mathbf{g} = \frac{\partial \max_{c \leq C} \mathcal{D}(\mathbf{z})[c]}{\partial \mathbf{z}} \in \mathbb{R}^{1 \times d}$$

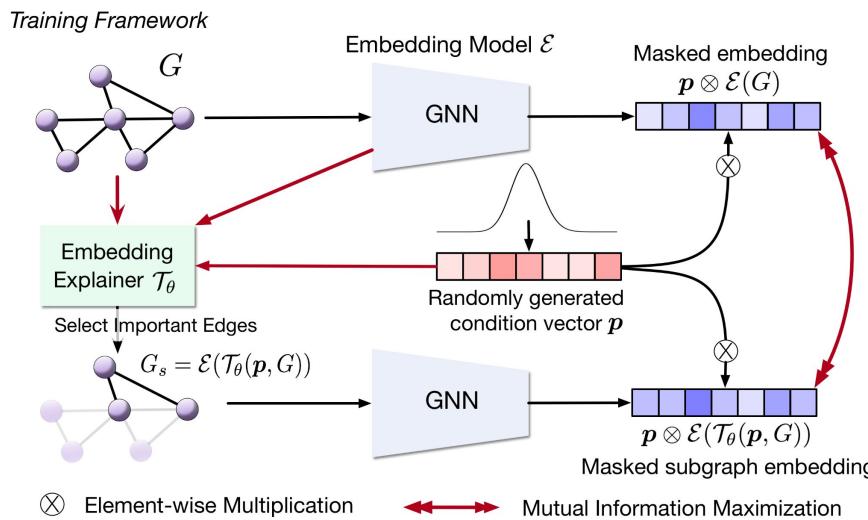
$$\mathbf{p} = \text{ReLU}(\text{norm}(\mathbf{g}^T))$$

- Value in a dimension with higher gradient indicates that the dimension is more important for a certain downstream task

# The Embedding Explainer & Training



- Inputs: the graph data, the GNN encoder, a condition vector.
- Output: the importance score for each edge.



*Objective:*

$$\max_{\theta} \mathbb{E}_p [\text{MI}(p \otimes \mathcal{E}(G), p \otimes \mathcal{E}(\mathcal{T}_\theta(p, G)))]$$

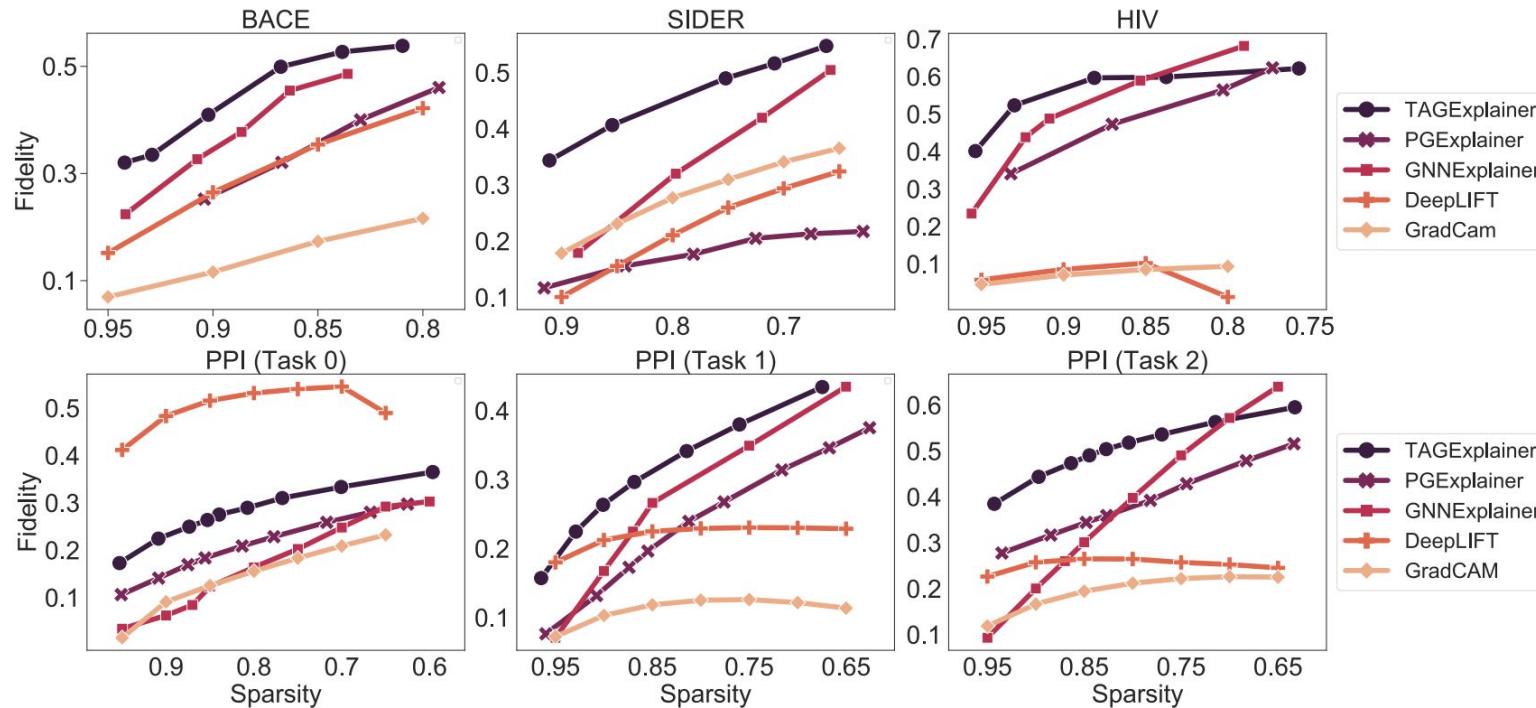
*Practically:*

$$\min_{\theta} -\frac{1}{N} \sum_{i=1}^N \left[ \log \frac{\exp\{(\mathbf{p} \otimes \mathbf{z}_i)^T (\mathbf{p} \otimes \mathbf{z}_{i,\theta})\}}{\sum_{j \neq i} \exp\{(\mathbf{p} \otimes \mathbf{z}_i)^T (\mathbf{p} \otimes \mathbf{z}_{j,\theta})\}} \right]$$

# Experimental Studies



TEXAS A&M UNIVERSITY  
Engineering



Xie et al., Task-agnostic Graph Explanations. NeurIPS'22.

# Experimental Studies



Table 6. Fidelity scores with controlled sparsity on the E-commerce product dataset. Each column corresponds to one explainer model trained on different tasks or without downstream task. Underlines highlight the best explanation quality in terms of fidelity, on the same level of sparsity.

| Eval on | PGExplainer (trained on) |                       |                       | TAGE<br>w/o downstream |
|---------|--------------------------|-----------------------|-----------------------|------------------------|
|         | Buyers                   | Sellers               | Reviews               |                        |
| Buyers  | <b>0.2009 ±0.2233</b>    | 0.1731 ±0.3774        | 0.1740 ±0.4463        | <b>0.2713 ±0.1834</b>  |
| Sellers | 0.5465 ±0.4773           | <b>0.3246 ±0.4026</b> | 0.1128 ±0.3019        | <b>0.6515 ±0.3426</b>  |
| Reviews | 0.4178 ±0.3683           | 0.1258 ±0.3492        | <b>0.2310 ±0.4178</b> | <b>0.5692 ±0.4214</b>  |

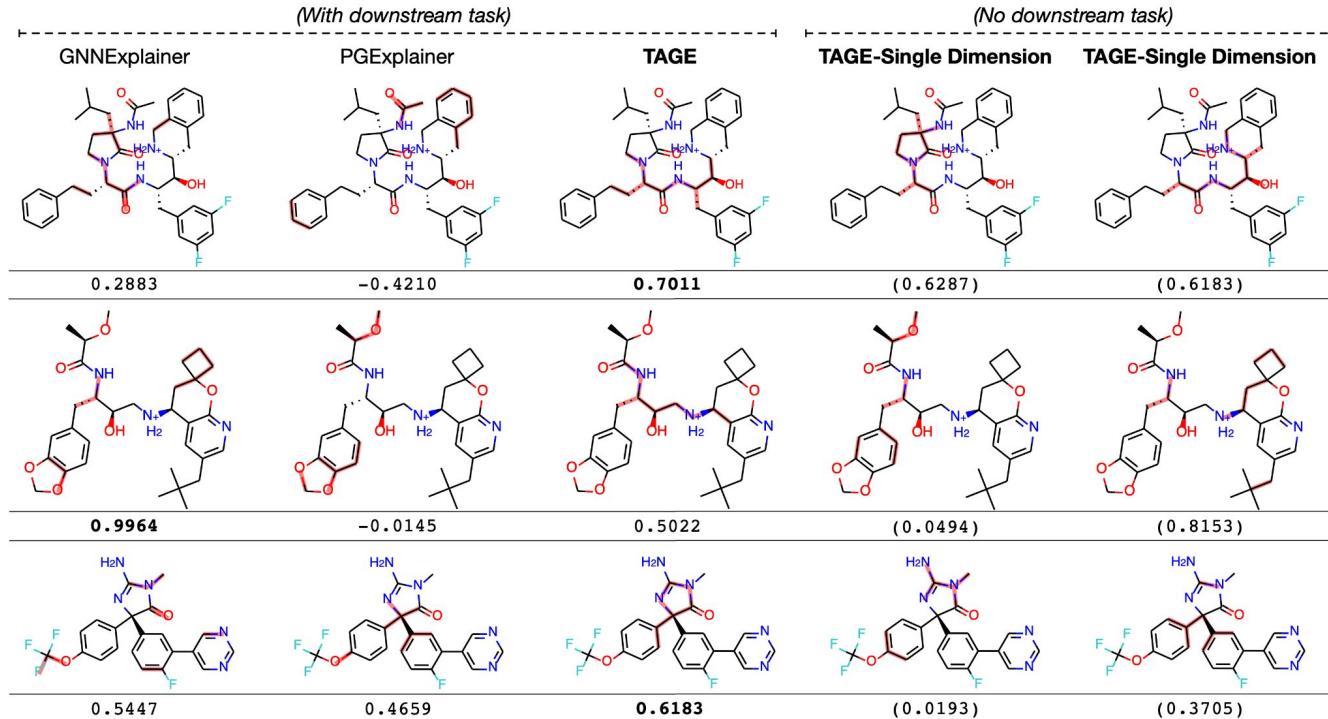
Table 3. Comparison of computational time cost among three learning-based GNN explainers on the PPI dataset. The left two columns record time cost breakdown for  $T$  downstream tasks. The fourth column estimates the total time cost for explaining all 121 tasks of PPI. The last row shows the speedup times compared to GNNExplainer and PGExplainer, respectively.

| Time cost    | Training (s)              | Inference (s)  | Total time (T=1) (s) | Est. total for 121 tasks |
|--------------|---------------------------|----------------|----------------------|--------------------------|
| GNNExplainer | $20040.1*T$               | —              | 20040.1              | 28 d                     |
| PGExplainer  | $7117.0*T$                | <b>427.2*T</b> | 7604.2               | 10.7 d                   |
| TAGE         | <b>1405.3</b>             | $582.7*T$      | <b>1988.0</b>        | <b>0.83 d</b>            |
| Speedup      | <b>14.3*T × / 5.1*T ×</b> | — / 0.73 ×     | <b>10.1× / 3.8×</b>  | <b>33.7× / 12.9×</b>     |

# Visualizations: Drug discovery



TEXAS A&M UNIVERSITY  
Engineering



Xie et al., Task-agnostic Graph Explanations. NeurIPS'22.



- Self-supervised image denoising
- Self-supervised graph representation learning
- Self-supervised explainer models
- Self-supervised learning for genetics studies

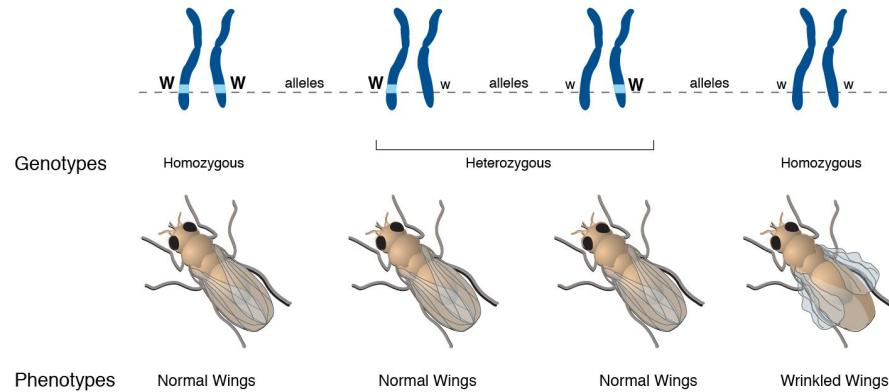
# Genome-Wide Association Studies



TEXAS A&M UNIVERSITY  
Engineering

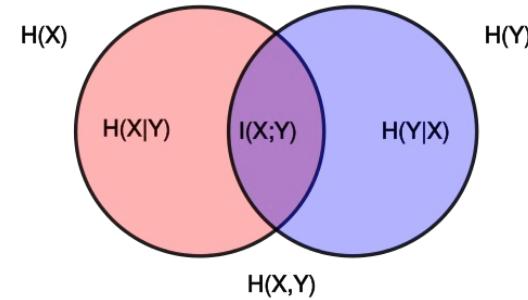
- Given traits, to identify SNPs that are associated with each trait.
- Our problem: learning representations from given high-dimensional data with neural networks. The representation should encode **as much information about the gene as possible**.

**Xie et al., Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies. Under review at NeurIPS'23.**



National Human Genome Research Institute.  
<https://www.genome.gov/>

- Given traits, to identify SNPs that are associated with each trait.
- Our problem: learning representations from given high-dimensional data with neural networks. The representation should encode **as much information about the gene as possible**.
  - Mutual Information (MI) maximization



*Xie et al., Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies. Under review at NeurIPS'23.*

- The mutual information between two random variables is defined as the KL-divergence between their joint distribution and the product of the marginals.
- MI Estimation is usually based on contrastive learning: to discriminate whether a pair of variables (e.g., MRI and genetic) belong to the same instance/individual.
- Jensen-Shannon Estimator:

$$\mathcal{L}_{\text{JSE}}(\mathbf{B}; \theta, \phi) = -\frac{1}{|B|} \sum_{(\mathbf{Y}, \mathbf{G}) \in \mathbf{B}} \log \left( \mathcal{D}_\phi(f_\theta(\mathbf{Y}), \mathbf{G}) \right) - \frac{1}{|B|(|B| - 1)} \sum_{(\mathbf{Y}, \mathbf{G}) \in \mathbf{B}} \left[ \sum_{(\mathbf{Y}', \mathbf{G}') \in \mathbf{B} \setminus \{(\mathbf{Y}, \mathbf{G})\}} \log \left( 1 - \mathcal{D}_\phi(f_\theta(\mathbf{Y}), \mathbf{G}') \right) \right],$$

*Xie et al., Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies. Under review at NeurIPS'23.*

- Contrastive learning are relatively well-studied for the visual representation learning (VRL) of natural images.
- What are the differences between VRL and the representation learning for GWAS?

# Difference 1: Goals of Learning Representations



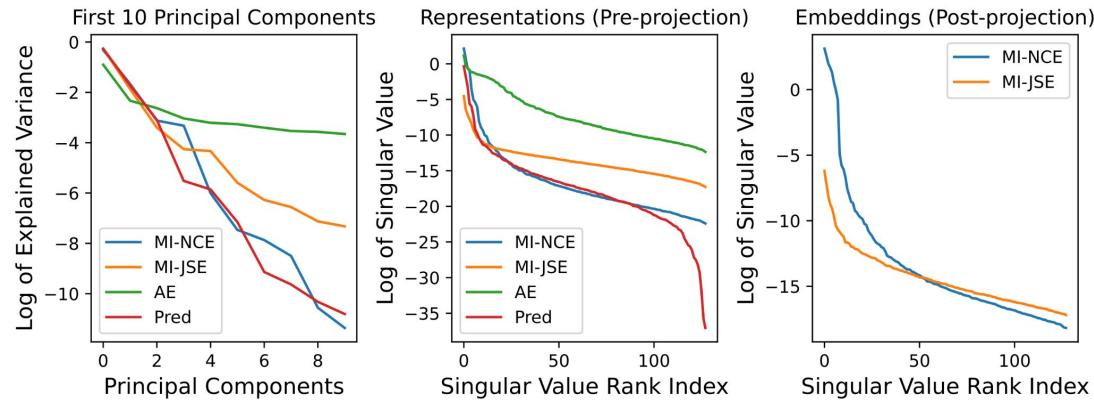
TEXAS A&M UNIVERSITY  
Engineering

- VRL: Aims to capture the key semantics or class information about main objects in images, not necessarily capturing all details in images.
- Representation for GWAS: No key semantics or class information. Need to capture as much details (related to gene) as possible.

*Xie et al., Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies. Under review at NeurIPS'23.*

# Challenge 1: Dimensional Collapse

- *Jing et al., 2021*: Typical contrastive approaches suffer from the dimensional collapse issue, where the learned representations occupy a lower-dimensional subspace than their designated dimensions.
- Trans-modal contrastive learning for GWAS suffers more from this issue.



*Xie et al., Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies. Under review at NeurIPS'23.*

## Difference 2: Augmentation v.s. Data Dimensions



TEXAS A&M UNIVERSITY  
Engineering

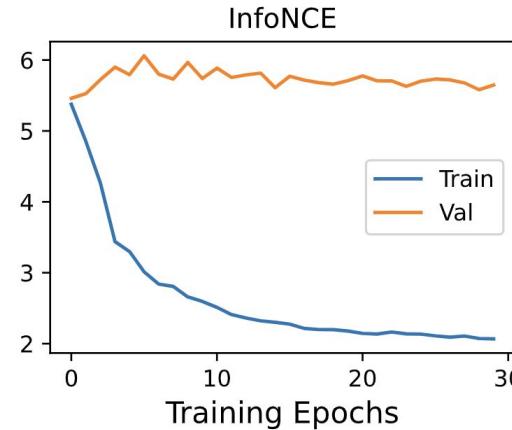
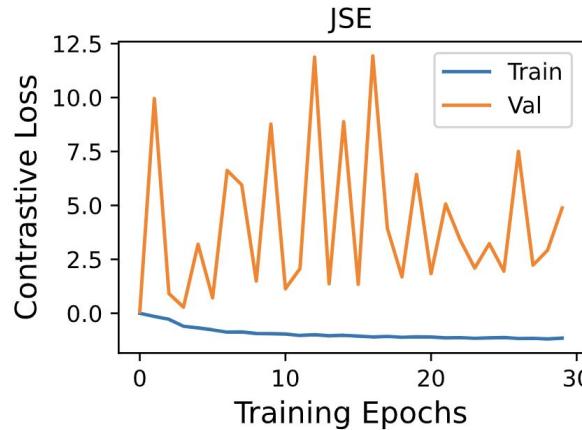
- Augmentations are crucial for contrastive learning.
- Augmentation approaches are limited for MRI data: many augmentations for natural images are unavailable.
- The high dimensional 3D MRI data makes it even worse.

*Xie et al., Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies. Under review at NeurIPS'23.*

## Challenge 2: Generalization Capability



- In our case, contrastive learning focuses too much on noise (non-generalizable associations) to identify individuals from MRI and genetic data.

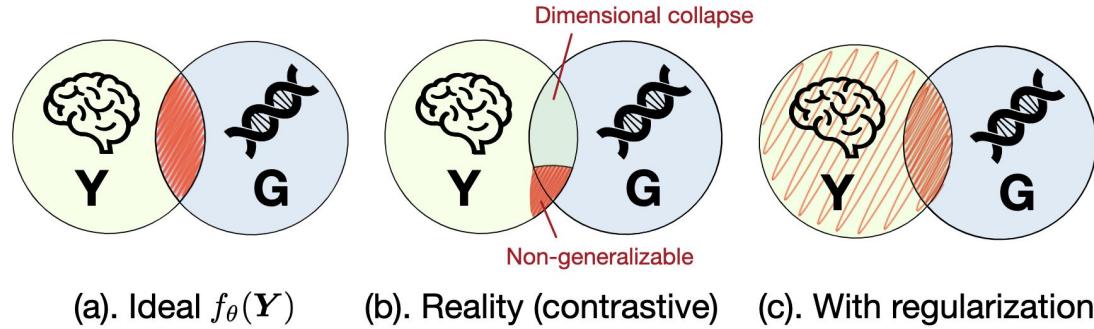


Xie et al., *Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies*. Under review at NeurIPS'23.

# MI Estimator with Regularizations



TEXAS A&M UNIVERSITY  
Engineering



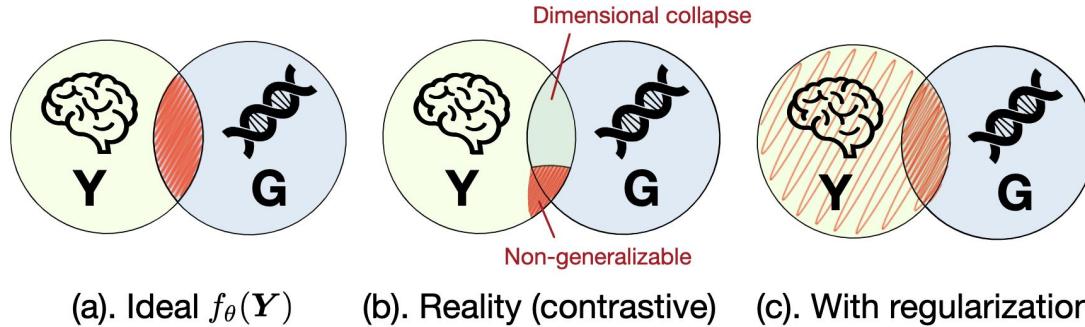
$$\max_{\theta} \left[ \hat{\mathcal{I}}(f_\theta(\mathbf{Y}), \mathbf{G}) + \lambda H(f_\theta(\mathbf{Y})) \right]$$

**Xie et al., Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies. Under review at NeurIPS'23.**

# MI Estimator with Regularizations



TEXAS A&M UNIVERSITY  
Engineering



$$\max_{\theta} \left[ \hat{\mathcal{I}}(f_\theta(\mathbf{Y}), \mathbf{G}) + \lambda \boxed{H(f_\theta(\mathbf{Y}))} \right] \rightarrow \mathcal{I}(f_\theta(\mathbf{Y}), \mathbf{Y}) + \cancel{H(f_\theta(\mathbf{Y})|\mathbf{Y})},$$

$$\mathcal{L}(\mathbf{B}; \theta, \phi, \psi) = \mathcal{L}_{\text{JSE}}(\mathbf{B}; \theta, \phi) + \frac{\lambda}{|\mathbf{B}|} \sum_{(\mathbf{Y}, \mathbf{G}) \in \mathbf{B}} \|\mathbf{Y} - h_\psi(f_\theta(\mathbf{Y}))\|^2$$

Xie et al., *Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies*. Under review at NeurIPS'23.

# What Else? To Better Encode Genetic Data



TEXAS A&M UNIVERSITY  
Engineering

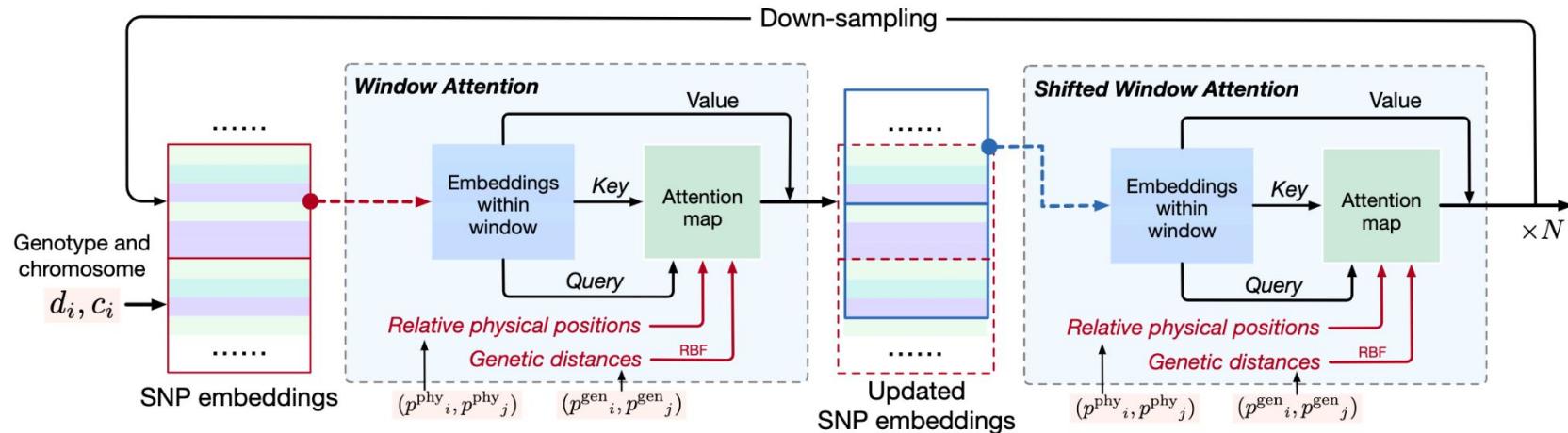
- The MI estimator also takes genetic data as input: We would like to encode as much information of genetic data as possible.
- Genetic Data is unstructured: Unable to use typical encoders such as convolutional layers.
- Size: > 650k SNPs. **Severe memory issue!**

*Xie et al., Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies. Under review at NeurIPS'23.*

# Genetics-Informed Transformer



- Significant optimized computational cost
- SNPs from arbitrary positions
- Flexibility to take any segments or subsets of the genetic data as inputs



Xie et al., *Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies*. Under review at NeurIPS'23.

# Experimental Results - GWAS

## T1 modality

|                                |                                       | # Independent Loci |           | MI (JSE)      | Heritability $h^2$                    |
|--------------------------------|---------------------------------------|--------------------|-----------|---------------|---------------------------------------|
|                                |                                       | All                | Unique    |               |                                       |
|                                | Random Init                           | 14                 | 1         | 1.2165        | $0.0756 \pm 0.0656$                   |
| <i>Predictive</i>              | Autoencoder (Patel et al., 2022)      | 26                 | 1         | 1.3120        | $0.3121 \pm 0.0769$                   |
|                                | Autoencoder-attention                 | 23                 | 4         | 1.3124        | $0.2984 \pm 0.0773$                   |
|                                | GenPred                               | 10                 | 0         | 1.2412        | $0.0918 \pm 0.1110$                   |
| <i>Contrastive</i>             | Barlow Twins (Zbontar et al., 2021)   | 11                 | 1         | 1.2996        | $0.0814 \pm 0.0636$                   |
|                                | SimCLR (Chen et al., 2020)            | 15                 | 1         | 1.2397        | $0.1448 \pm 0.1128$                   |
|                                | SimCLR-JSE                            | 17                 | 7         | 1.3044        | $0.1604 \pm 0.1151$                   |
| <i>Trans-Modal Contrastive</i> | InfoNCE (ContIG, Taleb et al. (2022)) | 11                 | 0         | 1.2299        | $0.1334 \pm 0.0588$                   |
|                                | Decorrelated InfoNCE                  | 13                 | 3         | 1.2382        | $0.0527 \pm 0.0349$                   |
|                                | GIM (Ours)                            | <b>40</b>          | <b>15</b> | <b>1.3681</b> | <b><math>0.3723 \pm 0.0305</math></b> |

## T2 modality

| Methods     | # All Indp Loci |
|-------------|-----------------|
| Autoencoder | 29              |
| Barlow-Twin | 7               |
| SimCLR      | 21              |
| SimCLR-JSE  | 15              |
| ContIG      | 22              |
| GIM (Ours)  | <b>38</b>       |

**Xie et al., Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies. Under review at NeurIPS'23.**

| Methods               | # Loci | Change    | # Unique Loci | Change   | $h^2$  | Most Significant New Gene |
|-----------------------|--------|-----------|---------------|----------|--------|---------------------------|
| Base-Contrastive      | 11     | -         | 0             | -        | 0.1334 | -                         |
| + Regularization      | 29     | +19 / -1  | 6             | +6 / -0  | 0.3390 | <i>CENPW</i>              |
| + Genetic Transformer | 32     | +14 / -9  | 10            | +7 / -3  | 0.3773 | <i>WNT16</i>              |
| + Random Cropping     | 36     | +17 / -13 | 13            | +10 / -7 | 0.3807 | <i>ITPR3</i>              |
| + Co-training         | 40     | +20 / -16 | 15            | +11 / -9 | 0.3723 | <i>MSRB3</i>              |

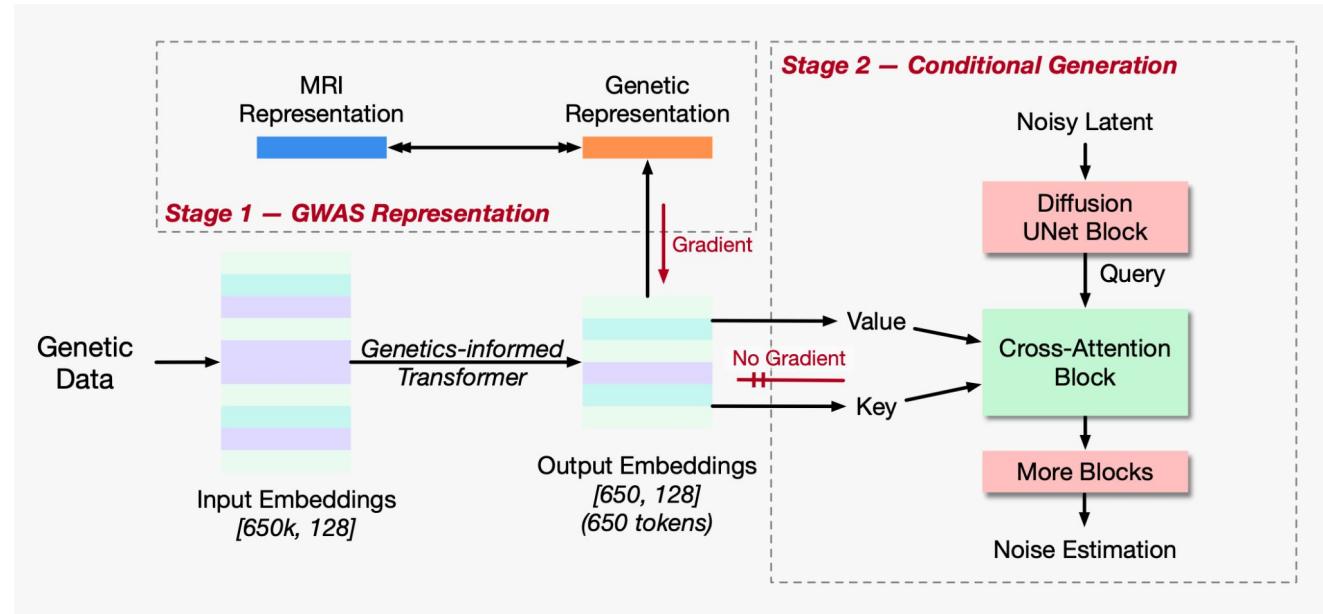
# Experimental Results - Conditional Generation



TEXAS A&M UNIVERSITY  
Engineering

Evaluating genetic encoder:

Use pre-trained genetic encoder to perform MRI generation conditional on gene.



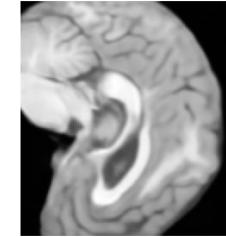
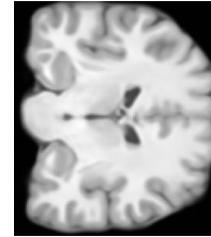
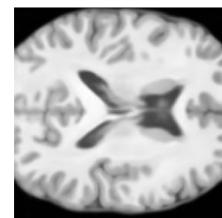
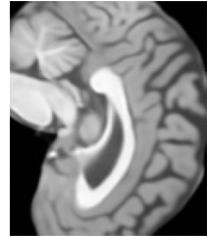
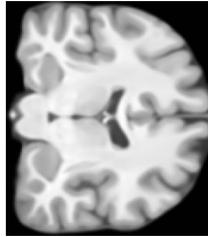
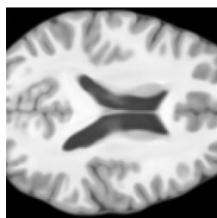
Xie et al., *Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies*. Under review at NeurIPS'23.

# Experimental Results - Conditional Generation

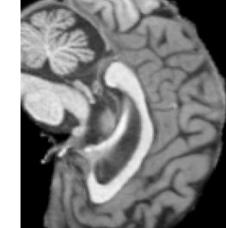
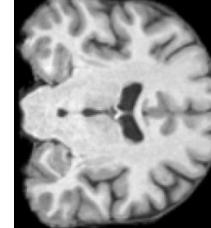
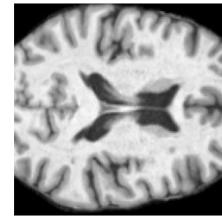
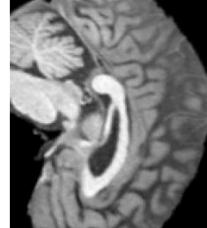
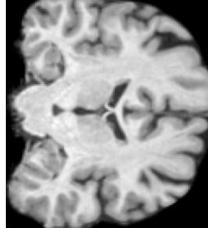
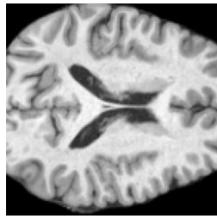


TEXAS A&M UNIVERSITY  
Engineering

Generated  
from gene:



Physically  
scanned:



Quantitative: **MSE** between  
generated and real MRI

| Models                    | h             | Downsampled | # Tokens | MSE           |
|---------------------------|---------------|-------------|----------|---------------|
| Avg. dist.                | —             | —           | —        | 0.3505        |
| Diffusion - Unconditional | —             | —           | —        | 0.2928        |
| Diffusion - Conditional   | Readout       | No          | 1        | 0.2858        |
|                           | Readout + MLP | No          | 1        | 0.2809        |
|                           | MLP           | Yes         | 60       | 0.2747        |
|                           | MLP           | No          | 650      | <b>0.2720</b> |

Xie et al., *Genetic InfoMax: Exploring Mutual Information Maximization in High-Dimensional Imaging Genetics Studies*. Under review at NeurIPS'23.

# Conclusions



TEXAS A&M UNIVERSITY  
Engineering

- My research work aims to improve the applicability and reliability of machine learning and deep learning approaches to drive broader industrial applications and scientific discovery (AI4Science).

- My research work aims to improve the applicability and reliability of machine learning and deep learning approaches to drive broader industrial applications and scientific discovery (AI4Science).
- Self-supervised
  - Denoising (as the starting point)
  - Graph representation learning (as a generalization)
  - Explanation (to enable reliability in industrial setting)
  - GWAS representation learning (to boost scientific discoveries)

**Yaochen Xie**, Sumeet Katariya, Xianfeng Tang, Edward Huang, Nikhil Rao, Karthik Subbian, Shuiwang Ji, “Task-Agnostic Graph Explanations.” The 36th Annual Conference on Neural Information Processing Systems (**NeurIPS**), 2022

**Yaochen Xie**\*, Zhao Xu\*, and Shuiwang Ji, “Self-Supervised Representation Learning via Latent Graph Prediction.” International Conference on Machine Learning (**ICML**), 24460-24477, 2022.

**Yaochen Xie**, Zhao Xu, Jingtun Zhang, Zhengyang Wang, and Shuiwang Ji, “Self-supervised Learning of Graph Neural Networks: A Unified Review.” IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), 2022.

**Yaochen Xie**, Yu Ding, Shuiwang Ji, “Augmented Equivariant Attention Networks for Microscopy Image Transformation.” IEEE Transactions on Medical Imaging (**TMI**), 2022.

**Yaochen Xie**, Zhengyang Wang, and Shuiwang Ji, “Noise2Same: Optimizing A Self-Supervised Bound for Image Denoising.” The 34th Neural Information Processing Systems (**NeurIPS**), 20320-20330, 2020.

# List of Publications



TEXAS A&M UNIVERSITY  
Engineering

Xinyi Xu, Cheng Deng, **Yaochen Xie**, Shuiwang Ji, "Group Contrastive Self-Supervised Learning on Graphs." IEEE Transactions on Pattern Analysis and Machine Intelligence (**TPAMI**), 2022.

Zhengyang Wang\*, Meng Liu\*, Youzhi Luo\*, Zhao Xu\*, **Yaochen Xie\***, Limei Wang\*, Lei Cai\*, Qi Qi, Zhuoning Yuan, Tianbao Yang, Shuiwang Ji, "Advanced graph and sequence neural networks for molecular property prediction and drug discovery" **Bioinformatics**, 38 (9), 2579-2586, 2022.

Liu, Meng\*, Youzhi Luo\*, Limei Wang\*, **Yaochen Xie\***, Hao Yuan\*, Shurui Gui, Haiyang Yu et al. "DIG: A Turnkey Library for Diving into Graph Deep Learning Research." Journal of Machine Learning Research (**JMLR**) 22 (2021): 1-9.

Zhengyang Wang\*, **Yaochen Xie\***, and Shuiwang Ji, "Global Voxel Transformer Networks for Augmented Microscopy." **Nature Machine Intelligence**, 3: 161-171, 2021.

Liu, Meng\*, Cong Fu\*, Xuan Zhang, Limei Wang, **Yaochen Xie**, Hao Yuan, Youzhi Luo, Zhao Xu, Shenglong Xu, and Shuiwang Ji. "Fast Quantum Property Prediction via Deeper 2D and 3D Graph Networks." AI for Science: Mind the Gaps workshop at **NeurIPS** 2021.

# Acknowledgement

Thank you to

- **Lin Gui**
- My advisor: **Dr. Shuiwang Ji**
- My committee members:
  - **Dr. Yu Ding**,
  - **Dr. Ruihong Huang**, and
  - **Dr. Bobak Mortazavi**
- All my collaborators, all DIVE members





TEXAS A&M UNIVERSITY  
**Engineering**

Thank you!