Learning Divergence Fields for Shift-Robust Graph Representations

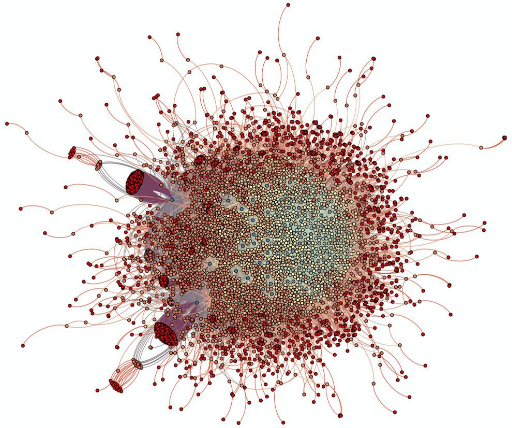
International Conference on Machine Learning (ICML), 2024

Qitian Wu, Fan Nie, Chenxiao Yang, Junchi Yan
Shanghai Jiao Tong University

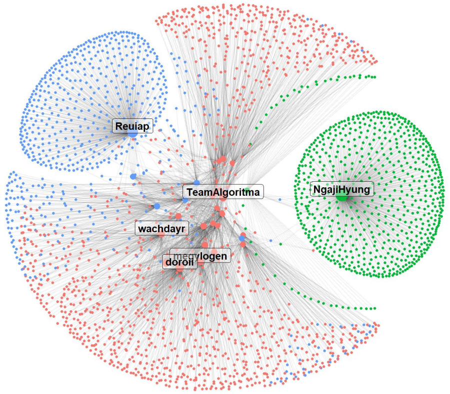
Paper: https://arxiv.org/pdf/2406.04963
Code: https://github.com/fannie1208/GLIND

Data with Explicit Structures

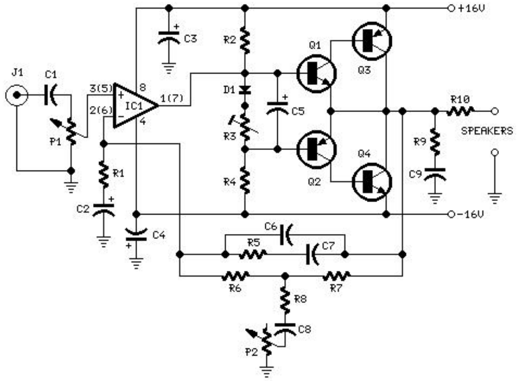
- Real-world data involves observed graph structures



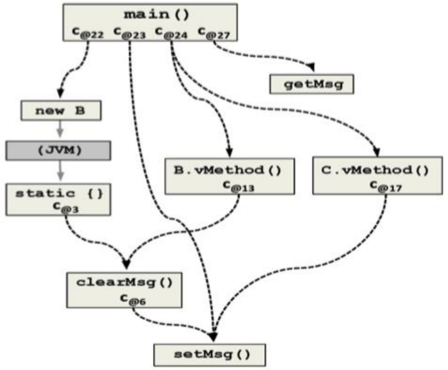
protein interactions



social networks



circuit graphs



code structures

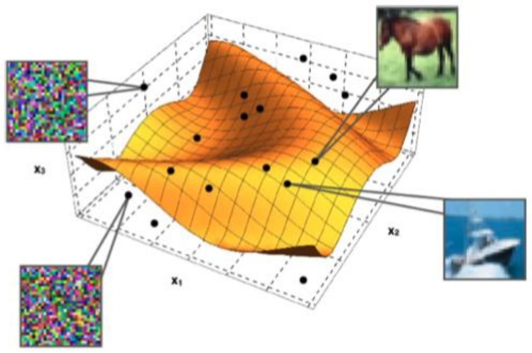
- **Characteristics of data with explict structures**

  1) Topological and geometric patterns (non-Euclidean space)

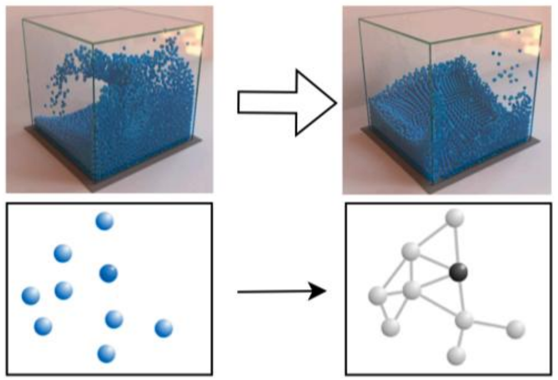
  2) Varying scales, sizes and properties

Data with Implicit Structures

❑ Real-world data involves unobserved graph structures



data manifold geometries
[Sebastian et al., 2021]



unknown physical interactions
[Alvaro et al., 2020]

infectious disease transmission
[Brockmann et al., 2013]

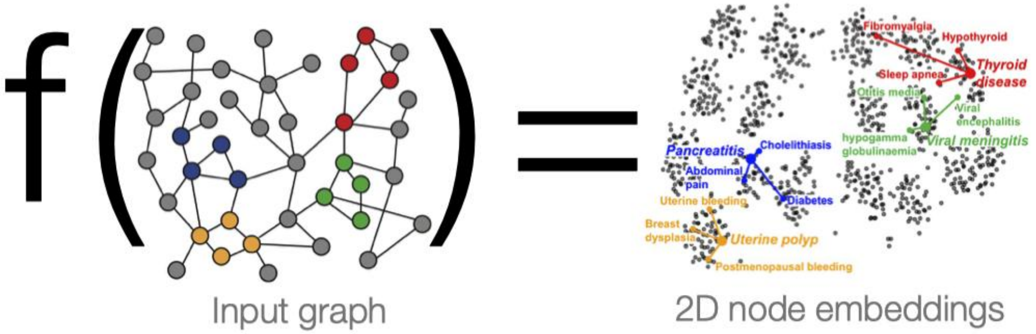
❑ Characteristics of data with implicit structures

1) Difficulty in inferring latent structures

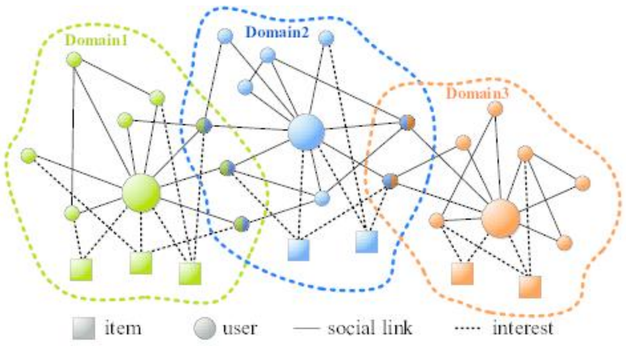
2) Scalability for large-scale systems

Graph Learning with Distribution Shifts

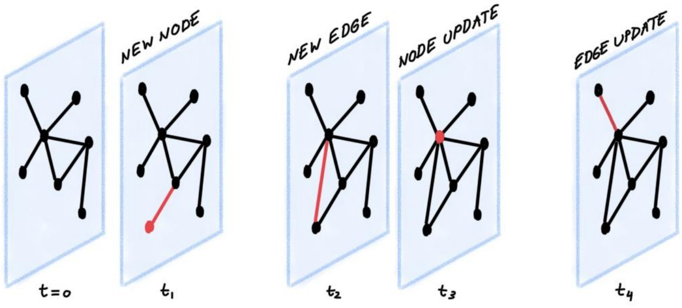
- **Graph representation learning:** find a functional map that converts nodes in a graph into embeddings in latent space



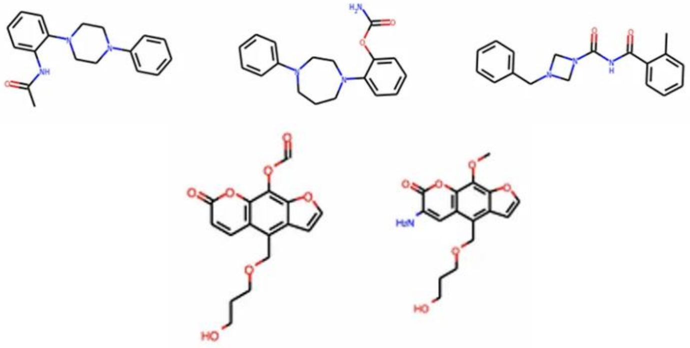
- **Graph distribution shifts:** difference between train and test data



Graphs from multiple domains



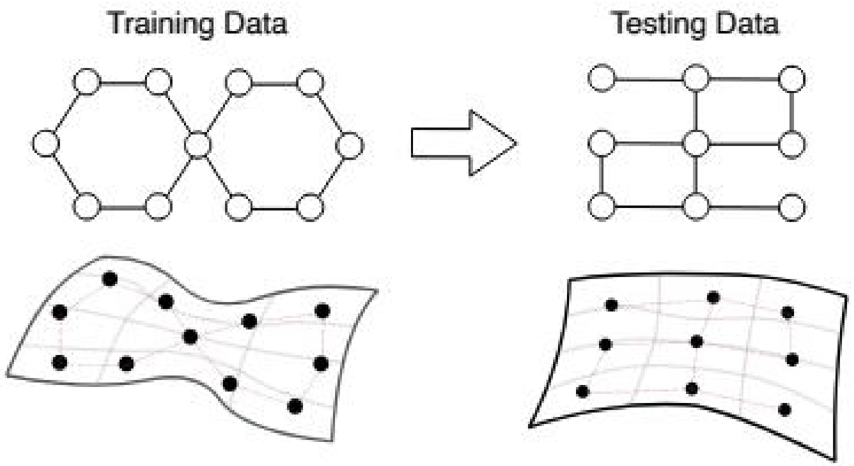
Temporal dynamic networks



Molecules with distinct drug likeness

Challenges of Distribution Shifts

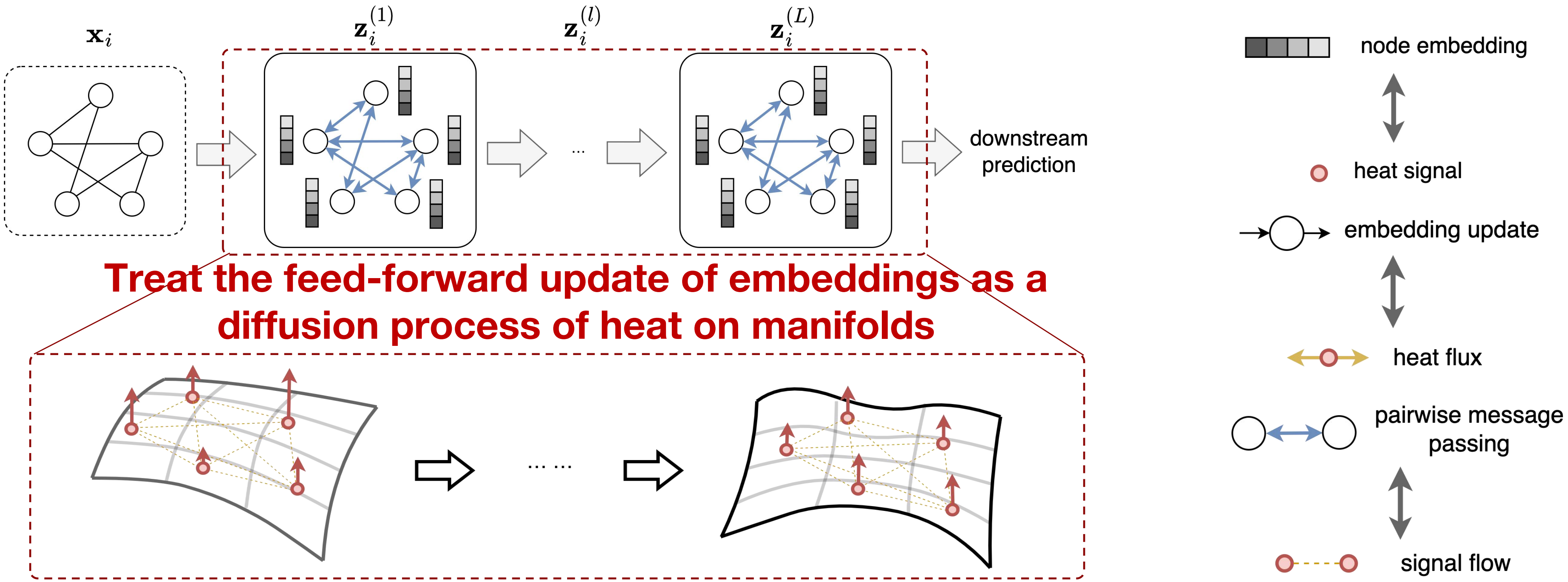
- **Generalization:** from training data to out-of-distribution testing data
  - Distribution shifts cause different data distributions $P_{train}(\mathcal{D}) \neq P_{test}(\mathcal{D})$
  - New data from unknown distribution are unseen by training
- **Latent geometry** behind observed data
  - Label of each instance depends on the instance itself and other instances
  - Interdependence of data points significantly increases the difficulty for generalization



How to model the generalizable predictive relations from inputs of interdependent data with certain geometries to their labels?

Message Passing as A Diffusion Process

❑ **Geometric diffusion**: a continuous process of **neural message passing**

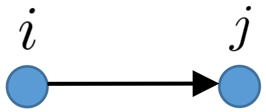


Qitian Wu et al., DIFFormer: Scalable (Graph) Transformers Induced by Energy Constrained Diffusion, ICLR 2023

Diffusion Equations on Graphs

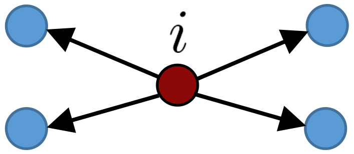
❑ **The diffusion process over N points driven by pairwise interactions:**

$$\frac{\partial z(u,t)}{\partial t} = \nabla^* \left(D(u,t) \odot \nabla z(u,t) \right), \quad z(u,0) = z_0(u), t \geq 0, u \in \Omega$$



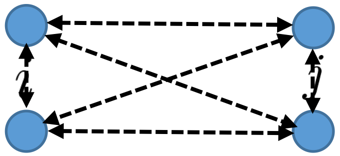
gradient

$$(\nabla \mathbf{Z}(t))_{uv} = \mathbf{z}_u(t) - \mathbf{z}_v(t)$$



divergence

$$(\nabla^*)_u = \sum_{v, a_{uv}=1} \mathbf{d}_v(\mathbf{Z}(t), u, t) \, (\nabla \mathbf{Z}(t))_{uv}$$



diffusivity function

$$\mathbf{d}(\mathbf{Z}(t), u, t)$$

❑ **Diffusion over discrete space of N nodes with latent structures:**

$$\frac{\partial \mathbf{z}_u(t)}{\partial t} = \sum_{v, a_{uv}=1} \mathbf{d}_v(\mathbf{Z}(t), u, t) \left(\mathbf{z}_v(t) - \mathbf{z}_u(t) \right), \quad \mathbf{Z}(0) = [\mathbf{x}_u]_{u=1}^N, t \geq 0$$

Qitian Wu et al., DIFFormer: Scalable (Graph) Transformers Induced by Energy Constrained Diffusion, ICLR 2023

Diffusion with Stochastic Diffusivity

- **Branching-structured divergence fields:** **the pairwise influence among data points could be driven by multiple criteria with uncertainty**

$$\frac{\partial \mathbf{z}_u(t)}{\partial t} = \boxed{\sum_{v, a_{uv}=1} d_{uv}^{(t)} \cdot (\mathbf{z}_v(t) - \mathbf{z}_u(t))}, \quad [d_{uv}^{(t)}]_{v=1}^{N} = \mathbf{d}_u^{(t)} \sim \boxed{p(\mathbf{d}^{(t)}|\mathbf{Z}(t), u, t)}$$

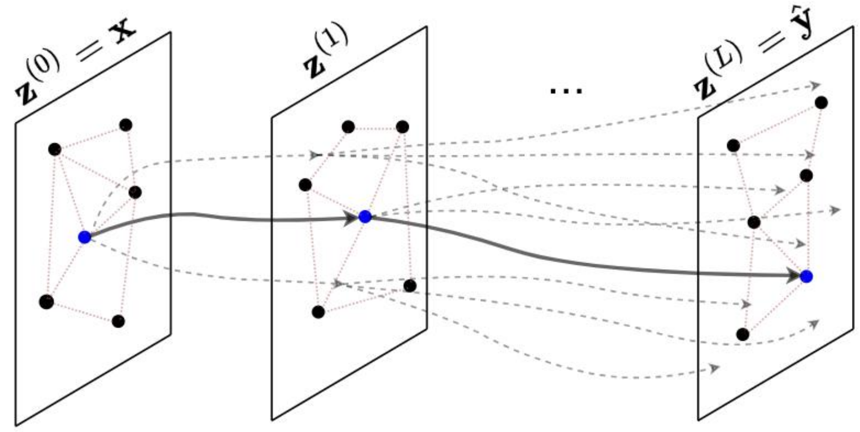
divergence: the amount of updated information

assume diffusivity to be generated from a probabilistic distribution

- **Diffusion trajectory:** **the discrete iterations induce layer-wise embeddings (** $\frac{\partial \mathbf{z}_u(t)}{\partial t} \approx \frac{\mathbf{z}_u^{(l+1)} - \mathbf{z}_u^{(l)}}{\tau}$ **)**

$$\mathbf{z}_u^{(l+1)} = \mathbf{z}_u^{(l)} + \alpha \sum_{v, a_{uv}=1} d_{uv}^{(l)} \cdot \left(\mathbf{z}_v^{(l)} - \mathbf{z}_u^{(l)}\right)$$

$$[d_{uv}^{(t)}]_{v=1}^{N} = \mathbf{d}_u^{(l)} \sim p(\mathbf{d}^{(l)}|\mathbf{z}_u^{(l)})$$



Probabilistic Formulation of Model

$$\boxed{\begin{aligned} \mathbf{z}_u^{(l+1)} &= \mathbf{z}_u^{(l)} + \alpha \sum_{v, a_{uv}=1} d_{uv}^{(l)} \cdot \left(\mathbf{z}_v^{(l)} - \mathbf{z}_u^{(l)}\right) \\ [d_{uv}^{(t)}]_{v=1}^{N} &= \mathbf{d}_u^{(l)} \sim p(\mathbf{d}^{(l)} | \mathbf{z}_u^{(l)}) \end{aligned}}$$

as a delta distribution

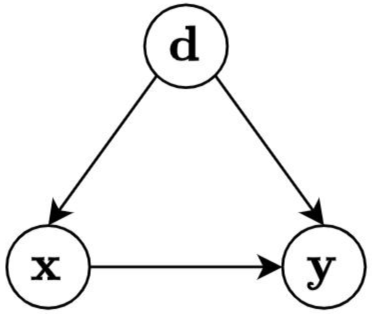
$$p_\theta(\mathbf{z}^{(l+1)} | \mathbf{z}^{(l)}, \mathbf{d}^{(l)}, \mathcal{G})$$

- ❑ **One step of model feedforward induces a predictive distribution:**

$$p_\theta(\mathbf{z}^{(l+1)} | \mathbf{z}^{(l)}, \mathcal{G}) = \mathbb{E}_{p(\mathbf{d}^{(l)} | \mathbf{z}^{(l)})}[p_\theta(\mathbf{z}^{(l+1)} | \mathbf{z}^{(l)}, \mathbf{d}^{(l)}, \mathcal{G})]$$

- ❑ **Likelihood of observed data for model training:**

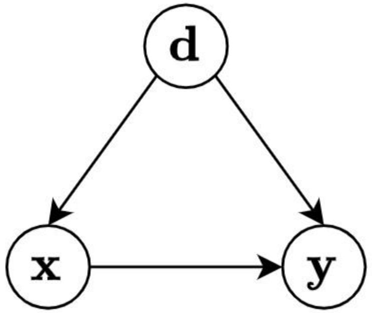
$$\begin{aligned} \log p_\theta(\mathbf{y} | \mathbf{x}, \mathcal{G}) &= \log \prod_{l=0}^{L-1} p_\theta(\mathbf{z}^{(l+1)} | \mathbf{z}^{(l)}, \mathcal{G}) \\ &= \sum_{l=0}^{L-1} \log \mathbb{E}_{p(\mathbf{d}^{(l)} | \mathbf{z}^{(l)})}[p_\theta(\mathbf{z}^{(l+1)} | \mathbf{z}^{(l)}, \mathbf{d}^{(l)}, \mathcal{G})] \end{aligned}$$



diffusivity is a latent confounder of x and y

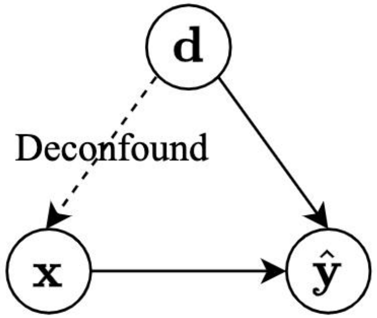
Deconfounded Learning/Causal Intervention

❑ Harmful effect: the confounding bias of latent diffusivity



- d establishes a shortcut (spurious correlation) between x and y
- Model training tends to exploit spurious correlation in training data
- Spurious correlation does not universally hold across environments

❑ Potential solution: cutting off the dependence between x and y



Key idea: replace $p_\theta(\mathbf{y}|\mathbf{x}, \mathcal{G})$ with $p_\theta(\mathbf{y}|do(\mathbf{x}), \mathcal{G})$

- According to Backdoor Adjustment in causal inference [Pearl et al., 2016]:

$$p_\theta(\mathbf{y}|do(\mathbf{x}), \mathcal{G}) = \sum_{\mathbf{d}} p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{d}, \mathcal{G}) p_0(\mathbf{d})$$

diffusivity is unobservable in real-world data sets

Deconfounded Learning/Causal Intervention

Theorem 1 (Variational Lower Bound of Causal Deconfounded Learning)

For any given diffusion model $p_\theta(\mathbf{z}^{(l+1)}|\mathbf{z}^{(l)}, \mathbf{d}^{(l)}, \mathcal{G})$, we have a lower bound of the deconfounded learning objective, i.e.,

$$\log p_\theta(\mathbf{y}|do(\mathbf{x}), \mathcal{G}) \geq \sum_{l=0}^{L-1} \mathbb{E}_{q_\phi(\mathbf{d}^{(l)}|\mathbf{z}^{(l)})} \left[\log p_\theta(\mathbf{z}^{(l+1)}|\mathbf{z}^{(l)}, \mathbf{d}^{(l)}, \mathcal{G}) \boxed{\frac{p_0(\mathbf{d}^{(l)})}{q_\phi(\mathbf{d}^{(l)}|\mathbf{z}^{(l)})}} \right]$$

a re-weighting term
penalize frequent
diffusivity components

In particular, the equality holds if and only if $q_\phi(\mathbf{d}^{(l)}|\mathbf{z}^{(l)}) = p_\theta(\mathbf{d}^{(l)}|\mathbf{z}^{(l)}, \mathbf{z}^{(l+1)}, \mathcal{G}) \cdot \frac{p_0(\mathbf{d}^{(l)})}{p(\mathbf{d}^{(l)}|\mathbf{z}^{(l)})}$.

Proof Sketch (see Appendix A in the papers):

- Backdoor adjustment $p_\theta(\mathbf{y}|do(\mathbf{x}), \mathcal{G}) = \sum_{\mathbf{d}^{(0)}, \cdots, \mathbf{d}^{(L-1)}} p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{d}^{(0)}, \cdots, \mathbf{d}^{(L-1)}, \mathcal{G}) p_0(\mathbf{d}^{(0)}, \cdots, \mathbf{d}^{(L-1)})$
- Variation lower bound

$$\sum_{l=0}^{L-2} \mathbb{E}_{q_\phi(\mathbf{d}^{(l)}|\mathbf{z}^{(l)})} \left[\log \sum_{\mathbf{z}^{(l+1)}} p_\theta(\mathbf{z}^{(l+1)}|\mathbf{z}^{(l)}, \mathbf{d}^{(l)}, \mathcal{G}) \frac{p_0(\mathbf{d}^{(l)})}{q_\phi(\mathbf{d}^{(l)}|\mathbf{z}^{(l)})} \right]$$
$$+ \mathbb{E}_{q_\phi(\mathbf{d}^{(L-1)}|\mathbf{z}^{(L-1)})} \left[\log p_\theta(\mathbf{z}^{(L)}|\mathbf{z}^{(L-1)}, \mathbf{d}^{(L-1)}, \mathcal{G}) \frac{p_0(\mathbf{d}^{(L-1)})}{q_\phi(\mathbf{d}^{(L-1)}|\mathbf{z}^{(L-1)})} \right],$$

Proposed Learning Objective

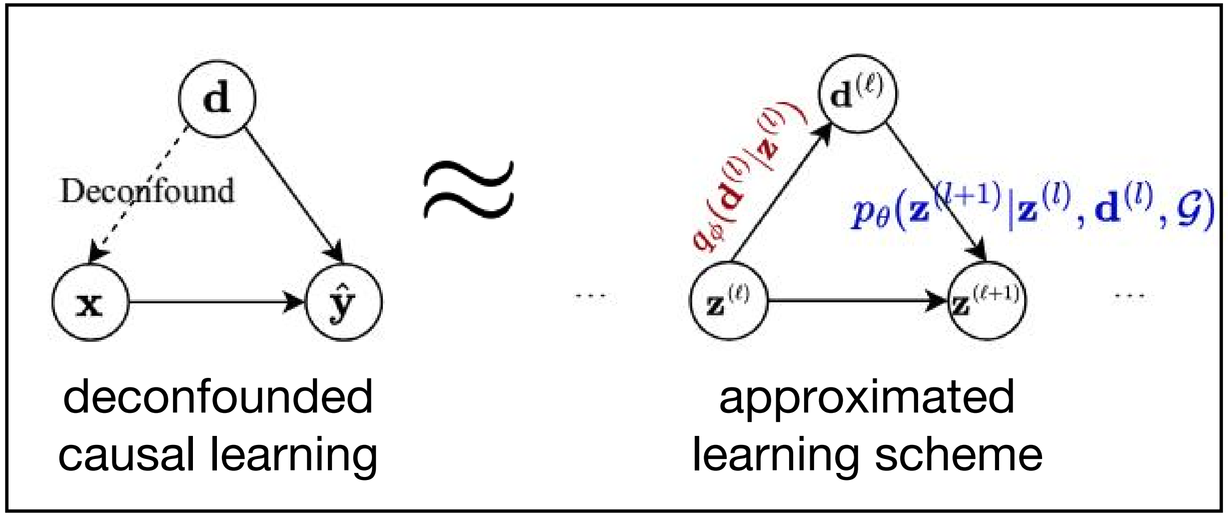
❑ **Learning objective: tractable lower bound of deconfounded learning**

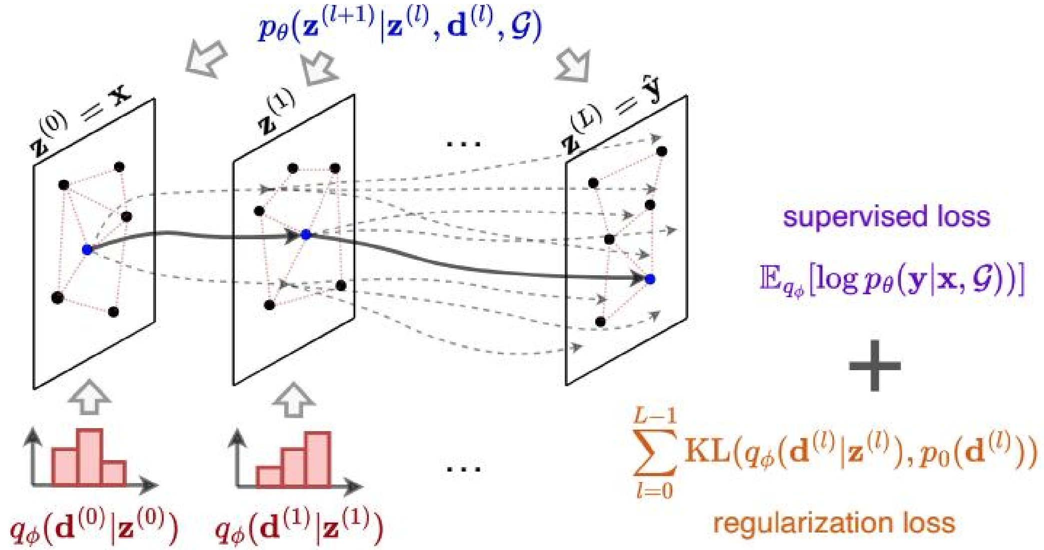
$$\mathbb{E}_{q_\phi(\mathbf{d}^{(0)}|\mathbf{z}^{(0)}),\cdots,q_\phi(\mathbf{d}^{(L-1)}|\mathbf{z}^{(L-1)})}\left[\log p_\theta(\mathbf{y}|\mathbf{x},\mathbf{d}^{(0)},\cdots,\mathbf{d}^{(L-1)},\mathcal{G})\right] - \sum_{l=0}^{L-1} \mathrm{KL}(q_\phi(\mathbf{d}^{(l)}|\mathbf{z}^{(l)}), p_0(\mathbf{d}^{(l)}))$$

estimate diffusivity at each layer

predict labels from inputs and estimated diffusivity

prior for diffusivity





Model Instantiation: Diffusivity Estimation

- **Latent diffusivity: assume diffusivity as samples from a set of hypothesis according to a multinomial distribution**

$$\mathbf{z}_u^{(l+1)} = \mathbf{z}_u^{(l)} + \sum_{k=1}^{K} h_{u,k}^{(l)} \sum_{v, a_{uv}=1} \boxed{d_{uv}^{(l,k)}} (\mathbf{z}_v^{(l)} - \mathbf{z}_u^{(l)}) \qquad \boxed{\mathbf{h}_u^{(l)}} \sim \mathcal{M}(\boldsymbol{\pi}_u^{(l)})$$

from a set of K diffusivity hypothesis $\{\mathbf{d}_u^{(l,k)}\}_{k=1}^K$

a one-hot vector from a multinomial dist.

- **Use Gumbel-Softmax to handle the non-differentiability of sampling:**

$$h_{u,k}^{(l)} = \frac{\exp\left(\left(\pi_u^{(l,k)} + g_k\right)/\tau\right)}{\sum_{k'} \exp((\pi_u^{(l,k')} + g_{k'})/\tau)}, \quad g_k \sim \mathrm{Gumbel}(0,1) \qquad [\pi_u^{(l,k)}]_{k=1}^K = \boldsymbol{\pi}_u^{(l)} = \mathrm{Softmax}(\mathbf{W}_L^{(l)} \mathbf{z}_u^{(l)})$$

- **Data-driven prior via mixture of posterior [Tomczak & Welling, 2018]:**

$$p_0(\mathbf{d}^{(l)}) = \frac{1}{T} \sum_{t=1}^{T} q(\mathbf{d}^{(l)} | \mathbf{z}^{(l)} = \boxed{\tilde{\mathbf{z}}_t^{(l)}})$$

embeddings of instances in the generated pseudo dataset $\{\tilde{\mathbf{x}}_t, \tilde{y}_t\}_{t=1}^T$ from a random graph model

Model Instantiation: Feedforward Propagation

❑ **Propagation layers:** assume diffusivity as different forms

- GLIND-GCN: Diffusivity as constant coupling matrix (graph adjacency)

$$\mathbf{z}_u^{(l+1)} = \mathbf{z}_u^{(l)} + \sum_{k=1}^{K} h_{u,k}^{(l)} \left(\sum_{v, a_{uv}=1} \frac{1}{\tilde{d}_u} \mathbf{W}_D^{(l,k)} \mathbf{z}_v^{(l)} + \mathbf{W}_S^{(l,k)} \mathbf{z}_u^{(l)} \right)$$

- GLIND-GAT: Diffusivity as time-dependent coupling matrix (graph attention)

$$\mathbf{z}_u^{(l+1)} = \mathbf{z}_u^{(l)} + \sum_{k=1}^{K} h_{u,k}^{(l)} \left(\sum_{v, a_{uv}=1} w_{uv}^{(l,k)} \mathbf{W}_D^{(l,k)} \mathbf{z}_v^{(l)} + \mathbf{W}_S^{(l,k)} \mathbf{z}_u^{(l)} \right) \qquad w_{uv}^{(l,k)} = \frac{\delta((\mathbf{c}^{(l,k)})^\top [\mathbf{W}_A^{(l,k)} \mathbf{z}_u^{(l)} \| \mathbf{W}_A^{(l,k)} \mathbf{z}_v^{(l)}])}{\sum_{w, a_{uw}=1} \delta((\mathbf{c}^{(l,k)})^\top [\mathbf{W}_A^{(l,k)} \mathbf{z}_u^{(l)} \| \mathbf{W}_A^{(l,k)} \mathbf{z}_w^{(l)}])}$$

- GLIND-Trans: Diffusivity as time-dependent coupling matrix (all-pair attention)

$$\mathbf{z}_u^{(l+1)} = \mathbf{z}_u^{(l)} + \sum_{k=1}^{K} h_{u,k}^{(l)} \left(\mathbf{W}_D^{(l,k)} \mathbf{b}_u^{(l,k)} + \mathbf{W}_S^{(l,k)} \mathbf{z}_u^{(l)} \right) \qquad \mathbf{b}_u^{(l,k)} = \sum_{v} \frac{\eta(\mathbf{W}_K^{(l,k)} \mathbf{z}_v^{(l)}, \mathbf{W}_Q^{(l,k)} \mathbf{k}_u^{(l)})}{\sum_{w=1}^{N} \eta(\mathbf{W}_K^{(l,k)} \mathbf{z}_w^{(l)}, \mathbf{W}_Q^{(l,k)} \mathbf{k}_u^{(l)})} \cdot \mathbf{z}_v^{(l)}$$

How to efficiently compute all-pair attention? DIFFormer [Wu et al., 2023]

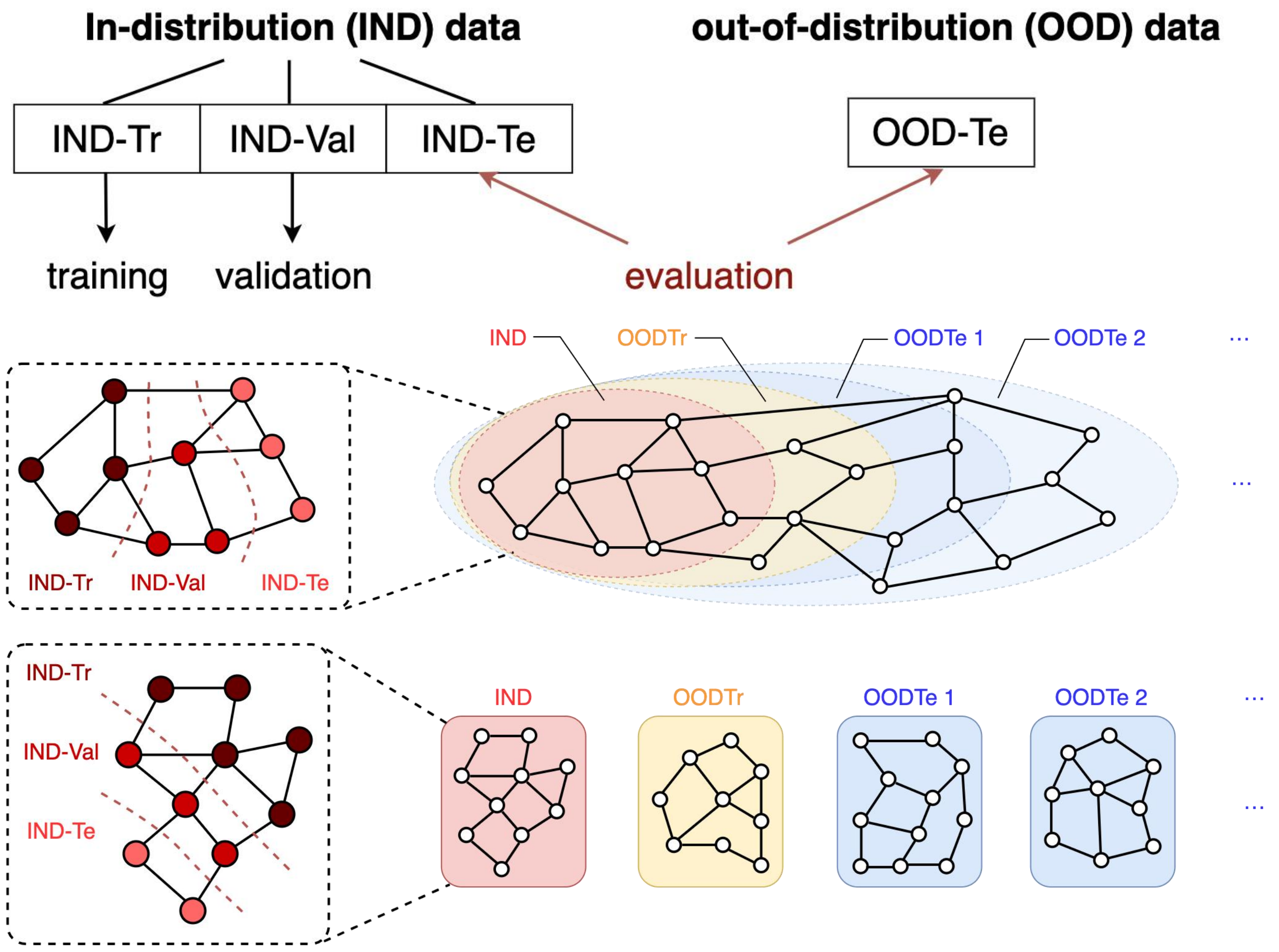
assume $\eta(\mathbf{a}, \mathbf{b}) = 1 + (\frac{\mathbf{a}}{\|\mathbf{a}\|_2})^\top \frac{\mathbf{b}}{\|\mathbf{b}\|_2}$

$$\mathbf{b}_u^{(l,k)} = \frac{\sum_{v=1}^{N} \mathbf{z}_v^{(l)} + \left(\sum_{v=1}^{N} (\mathbf{k}_v^{(l)})(\mathbf{z}_v^{(l)})^\top \right) (\mathbf{q}_u^{(l)})}{N + (\mathbf{q}_u^{(l)})^\top (\sum_{v=1}^{N} \mathbf{k}_v^{(l)})}$$

only require $O(N)$ for updating N instances

Experiment Protocols

- Split data into in-distribution and out-of-distribution portions; for IND data, randomly split into IND-Tr/IND-Val/IND-Te
- For temporal graph dataset: use time information for data split of IND and OOD
- For multi-graph dataset: use domain information for data split of IND and OOD



Qitian Wu, et al., Handling Distribution Shifts on Graphs: An Invariance Perspective, ICLR 2022

Qitian Wu, et al., Energy-based Out-of-Distribution Detection for Graph Neural Networks, ICLR 2023

Experiment Results

Testing results (Accuracy for *Arxiv*, ROC-AUC for *Twitch*) on real-world datasets

Method	Arxiv			Twitch		
	2014-2016	2016-2018	2018-2020	ES	FR	EN
ERM-GCN	56.33 ± 0.17	53.53 ± 0.44	45.83 ± 0.47	66.07 ± 0.14	52.62 ± 0.01	63.15 ± 0.08
IRM-GCN	55.92 ± 0.24	53.25 ± 0.49	45.66 ± 0.83	66.95 ± 0.27	52.53 ± 0.02	62.91 ± 0.08
GroupDRO-GCN	56.52 ± 0.27	53.40 ± 0.29	45.76 ± 0.59	66.82 ± 0.26	52.69 ± 0.02	62.95 ± 0.11
DANN-GCN	56.35 ± 0.11	53.81 ± 0.33	45.89 ± 0.37	66.15 ± 0.13	52.66 ± 0.02	63.20 ± 0.06
Mixup-GCN	56.67 ± 0.46	54.02 ± 0.51	46.09 ± 0.58	65.76 ± 0.30	52.78 ± 0.04	63.15 ± 0.08
EERM-GCN	-	-	-	67.50 ± 0.74	51.88 ± 0.07	62.56 ± 0.02
GLIND-GCN	**59.42 ± 0.33**	**56.84 ± 0.54**	**57.06 ± 1.21**	**67.72 ± 0.10**	**53.16 ± 0.08**	**64.18 ± 0.03**
ERM-GAT	57.15 ± 0.25	55.07 ± 0.58	46.22 ± 0.82	65.67 ± 0.02	52.00 ± 0.10	61.85 ± 0.05
IRM-GAT	56.55 ± 0.18	54.53 ± 0.32	46.01 ± 0.33	67.27 ± 0.19	52.85 ± 0.15	62.40 ± 0.24
GroupDRO-GAT	56.69 ± 0.27	54.51 ± 0.49	46.00 ± 0.59	67.41 ± 0.04	52.99 ± 0.08	62.29 ± 0.03
DANN-GAT	57.23 ± 0.18	55.13 ± 0.46	46.61 ± 0.57	66.59 ± 0.38	52.88 ± 0.12	62.47 ± 0.32
Mixup-GAT	57.17 ± 0.33	55.33 ± 0.37	47.17 ± 0.84	65.58 ± 0.13	52.04 ± 0.04	61.75 ± 0.13
EERM-GAT	-	-	-	66.80 ± 0.46	52.39 ± 0.20	62.07 ± 0.68
GLIND-GAT	**60.36 ± 0.36**	**58.98 ± 0.43**	**59.71 ± 0.53**	**67.82 ± 0.10**	**54.50 ± 0.12**	**64.32 ± 0.12**

Experiment Results

Testing RMSE for protein interaction dataset on different domains

Method	Hazbun	Krogan (LCMS)	Krogan (MALDI)	Lambert	Tarassov	Uetz	Yu
ERM-Trans	1.82 ± 0.17	1.63 ± 0.04	1.57 ± 0.03	1.49 ± 0.07	1.62 ± 0.03	1.52 ± 0.04	1.51 ± 0.04
IRM-Trans	1.66 ± 0.14	1.86 ± 0.04	1.84 ± 0.04	1.52 ± 0.07	1.76 ± 0.03	1.66 ± 0.05	1.66 ± 0.04
DANN-Trans	1.69 ± 0.11	1.66 ± 0.02	1.62 ± 0.03	1.39 ± 0.05	1.63 ± 0.01	1.49 ± 0.01	1.50 ± 0.01
GroupDRO-Trans	1.65 ± 0.13	1.68 ± 0.02	1.65 ± 0.02	1.48 ± 0.03	1.72 ± 0.01	1.53 ± 0.04	1.53 ± 0.01
Mixup-Trans	1.46 ± 0.13	1.79 ± 0.05	1.76 ± 0.04	1.50 ± 0.06	1.70 ± 0.05	1.56 ± 0.06	1.59 ± 0.06
EERM-Trans	1.68 ± 0.47	1.91 ± 0.23	1.92 ± 0.09	1.47 ± 0.05	1.79 ± 0.11	1.67 ± 0.07	1.65 ± 0.08
GLIND-TRANS	1.02 ± 0.07	1.38 ± 0.07	1.33 ± 0.05	1.08 ± 0.04	1.40 ± 0.04	1.20 ± 0.04	1.20 ± 0.04

- DDPIN (dynamic protein interaction dataset) contains multiple dynamic graphs
- Each dynamic graph is from a protein identification method
- Each node has a scalar-valued signal evolving with time and affecting the graph structure (co-expressed levels between proteins)

Experiment Results

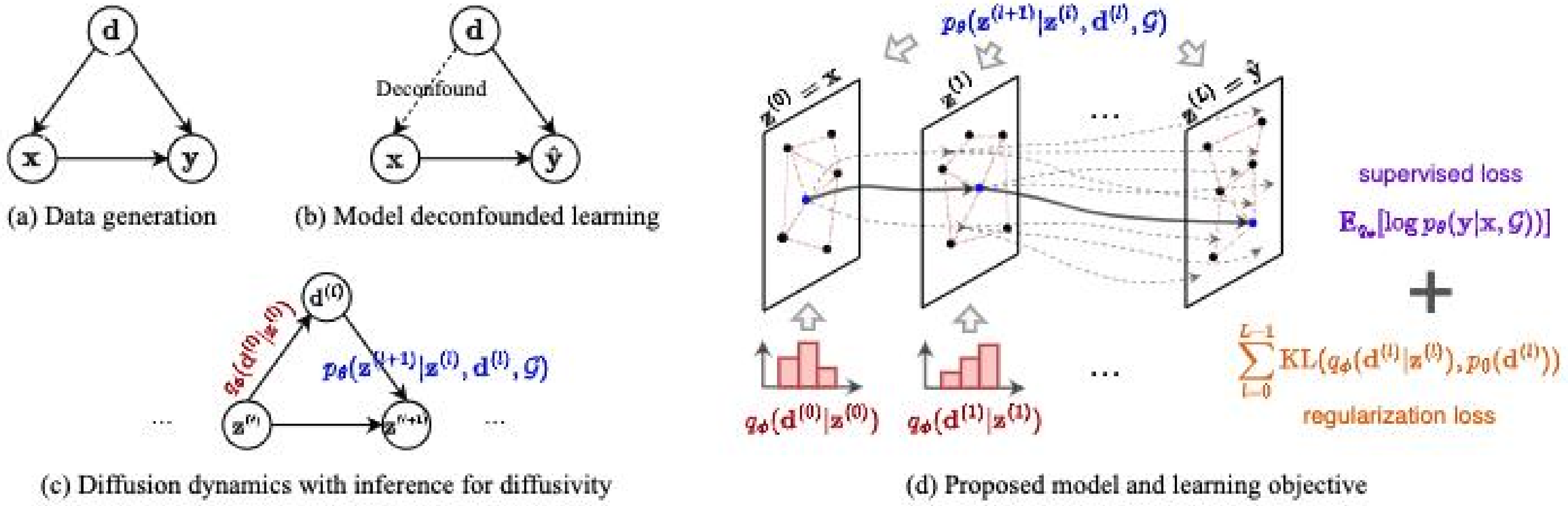
Testing Accuracy (%) for CIFAR and STL on different domains

Method	CIFAR			STL		
	150°	160°	170°	$k=8$	$k=9$	$k=10$
ERM-Trans	76.88 ± 0.11	77.51 ± 0.25	76.35 ± 0.28	76.53 ± 0.25	77.10 ± 0.65	77.90 ± 0.22
IRM-Trans	76.53 ± 0.03	77.11 ± 0.05	76.42 ± 0.31	76.95 ± 0.14	77.49 ± 0.25	78.02 ± 0.35
GroupDRO-Trans	76.94 ± 0.65	76.99 ± 0.31	76.37 ± 0.53	77.81 ± 0.59	78.01 ± 0.54	78.10 ± 0.27
DANN-Trans	76.91 ± 0.17	77.13 ± 0.37	76.61 ± 0.30	77.64 ±0.13	78.29 ± 0.54	78.19 ± 0.35
Mixup-Trans	77.49 ± 0.39	77.91 ± 0.14	77.45 ± 0.34	77.76 ± 0.30	78.32 ± 0.57	**78.73 ± 0.76**
EERM-Trans	79.68 ± 0.51	79.89 ± 0.32	78.82 ± 0.54	77.92 ± 0.93	78.58 ± 0.20	78.18 ± 0.38
GLIND-TRANS	**80.72 ± 0.39**	**81.06 ± 0.32**	**80.24 ± 0.38**	**78.06 ± 0.46**	**79.39 ± 0.28**	**78.41 ± 0.57**

- **Each instance is an image/text without observed interdependent structures**
- **Use k-nearest-neighbor to create a synthetic graph structure among instances**
- **Use different values of k and similarity functions (added with rotation angles) to introduce distribution shifts between training and test data**

Conclusion

We explore a geometric diffusion framework empowered by causal learning for shift-robust graph representations (out-of-distribution generalization)



Qitian Wu, et al., Handling Distribution Shifts on Graphs: An Invariance Perspective, ICLR 2022

Qitian Wu, et al., Energy-based Out-of-Distribution Detection for Graph Neural Networks, ICLR 2023

Qitian Wu, et al., DIFFormer: Scalable (Graph) Transformers Induced by Energy Constrained Diffusion, ICLR 2023