Université de Montréal

Mila

Molecular Representation Learning with Limited Data

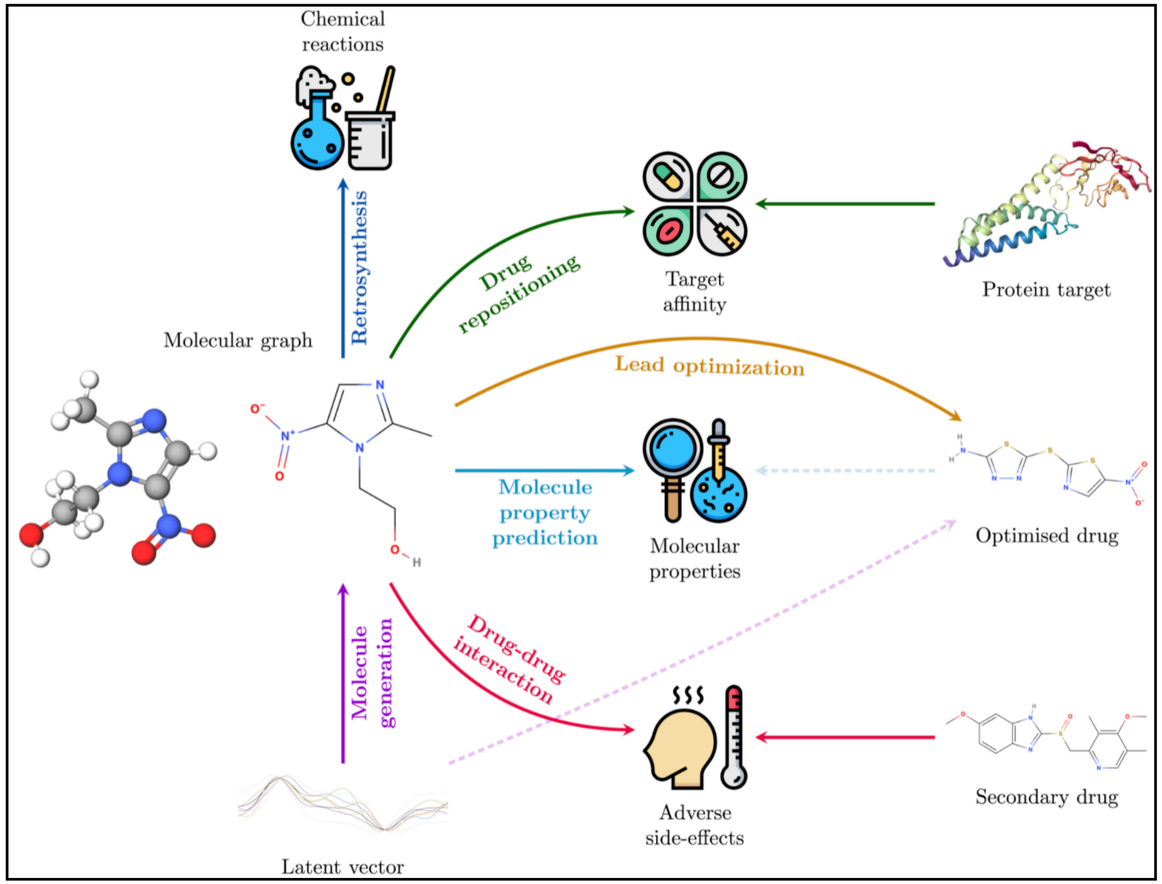
Shengchao Liu, Dec 2021

Pipeline

1. **Introduction**
2. **Motivation**
3. **Self-Supervised Learning (SSL) for Molecular Representation Learning**
4. **Multi-Task Learning (MTL) for Molecular Representation Learning**
5. **Conclusions and Future Directions**

1. Introduction

- Report: A Tough Road: Cost To Develop One New Drug Is $2.6 Billion:
  - Average length: 10-12 years, average cost: ~US $2.6 billion.
- Solution: AI-guided drug discovery:
  - Ongoing and promising.
  - Can accelerate multiple stages.



Credit to Liu, Shengchao, Deac, Andreea, and Tang, Jian.
“Graph Representation Learning for Drug Discovery." *Manuscript.*

1. Introduction

- A Tough Road: Cost To Develop One New Drug Is $2.6 Billion:
  - Average length: 10-12 years, average cost: ~US $2.6 billion.
- Solution: AI-guided drug discovery:
  - Ongoing and promising.
  - Can accelerate multiple stages.
  - Fundamental Challenges:
    - Molecular representation
    - Low-data
    - Class/Label imbalance
    - …

1. Introduction

- A Tough Road: Cost To Develop One New Drug Is $2.6 Billion:
  - Average length: 10-12 years, average cost: ~US $2.6 billion.
- Solution: AI-guided drug discovery:
  - Ongoing and promising.
  - Can accelerate multiple stages.
  - Fundamental Challenges:
    - Molecular representation
    - Low-data
    - Class/Label imbalance
    - …

2. Motivation

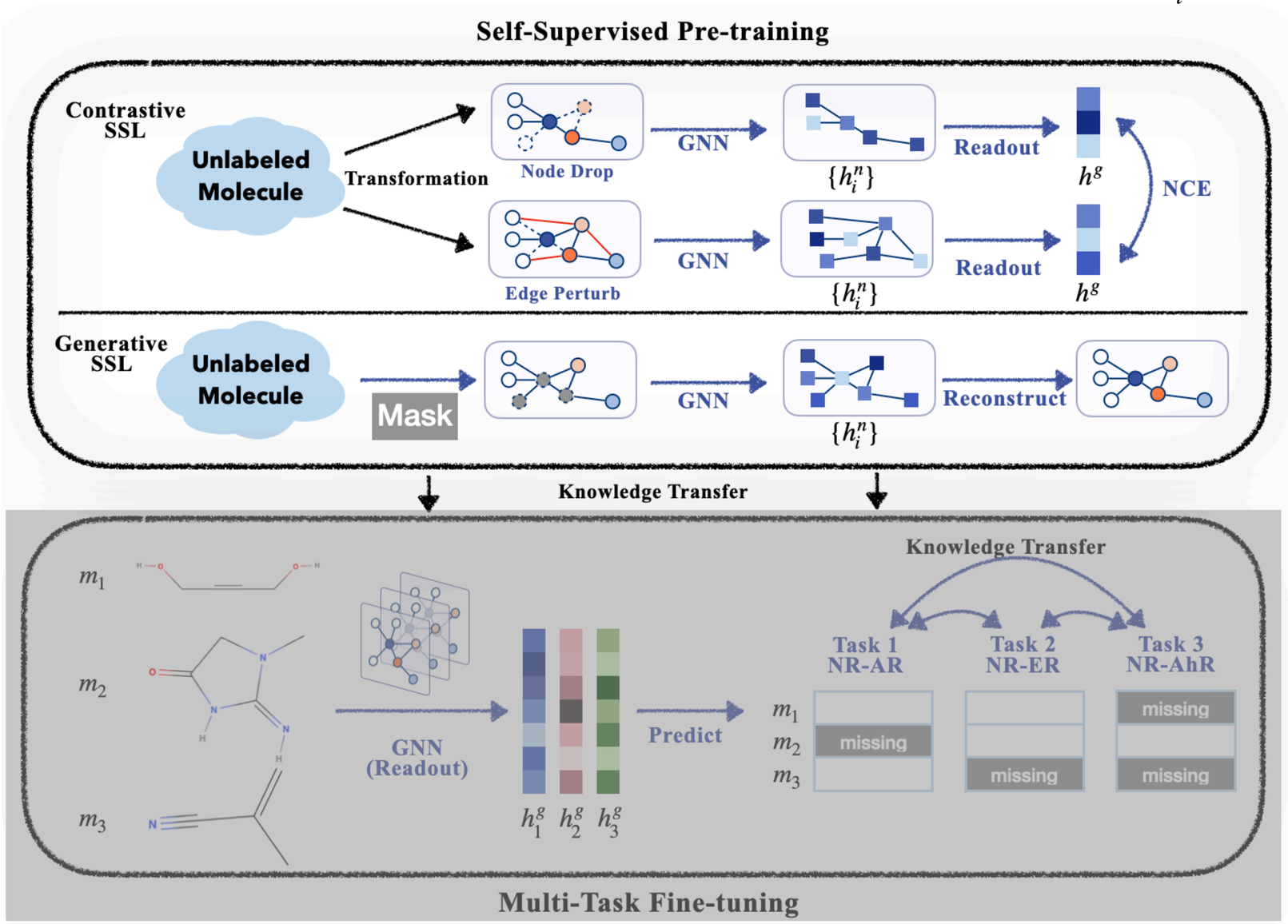
Q: How to handle such low-data issue?

A: Transferring knowledge.

- Self-supervised (unsupervised) pre-training:
  - Pre-training: patterns (CV), semantics (NLP), structures (Graph).
  - Fine-tuning: smaller dataset.
- Multi-task learning:
  - Joint learning.
  - Improved general performance.

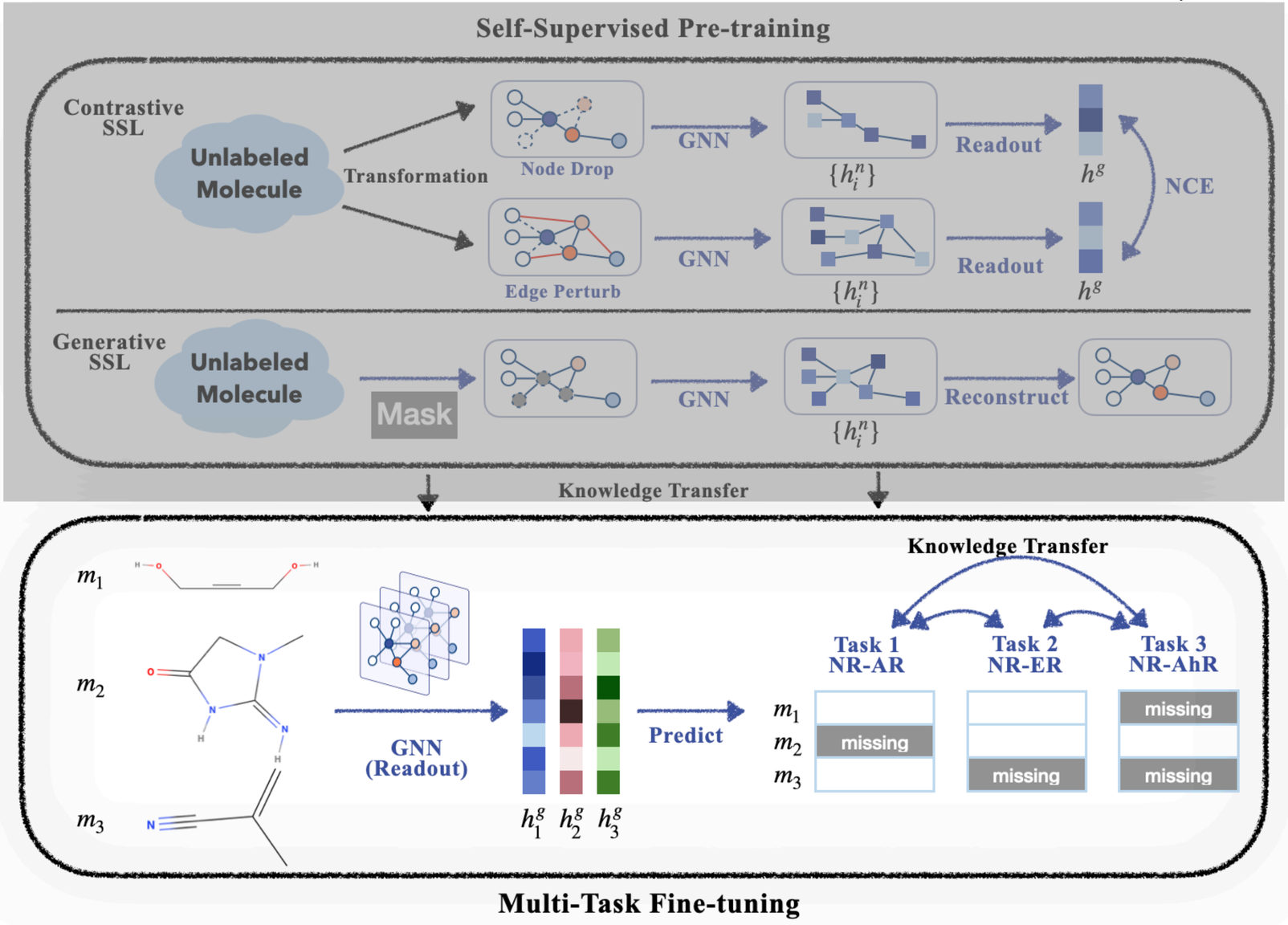
2. Motivation

m_i: molecule
h_i^n: node representation
h^g/h_i^g: graph representation



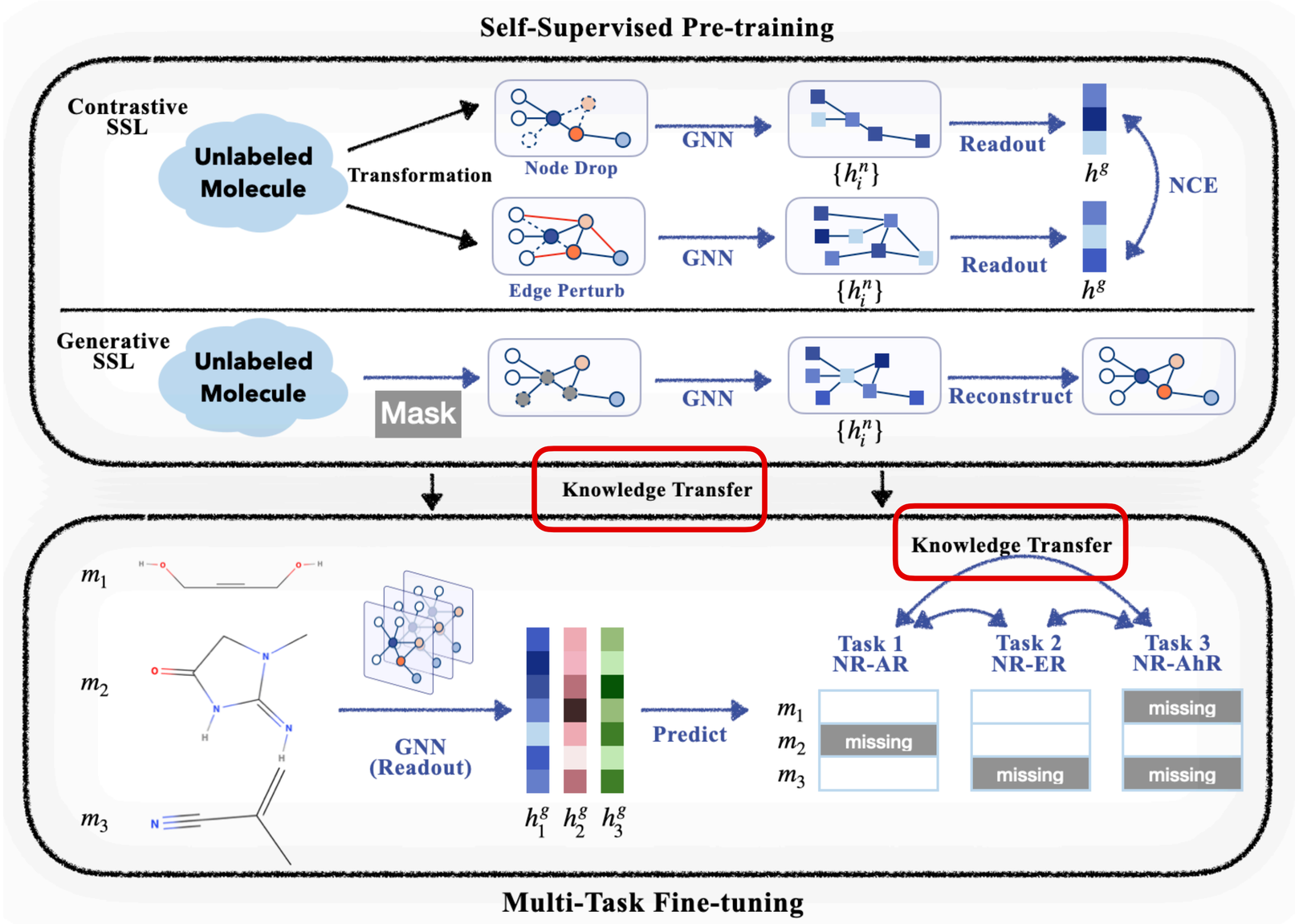
2. Motivation

m_i: molecule
h_i^n: node representation
h^g/h_i^g: graph representation



2. Motivation

Q: Can we better utilize the domain knowledge for transferring?



3. SSL for Molecular Property Prediction

3.1 Problem Definition

3.2 Motivation

3.3 Related Work

3.4 Preliminaries

3.5 Method: GraphMVP

3.6 Experiments

3.1 Problem Definition

Ultimate goal:

- Molecular property prediction on target (downstream) tasks.
- MoleculeNet [1]: only 2D topological information for molecular graph is available.

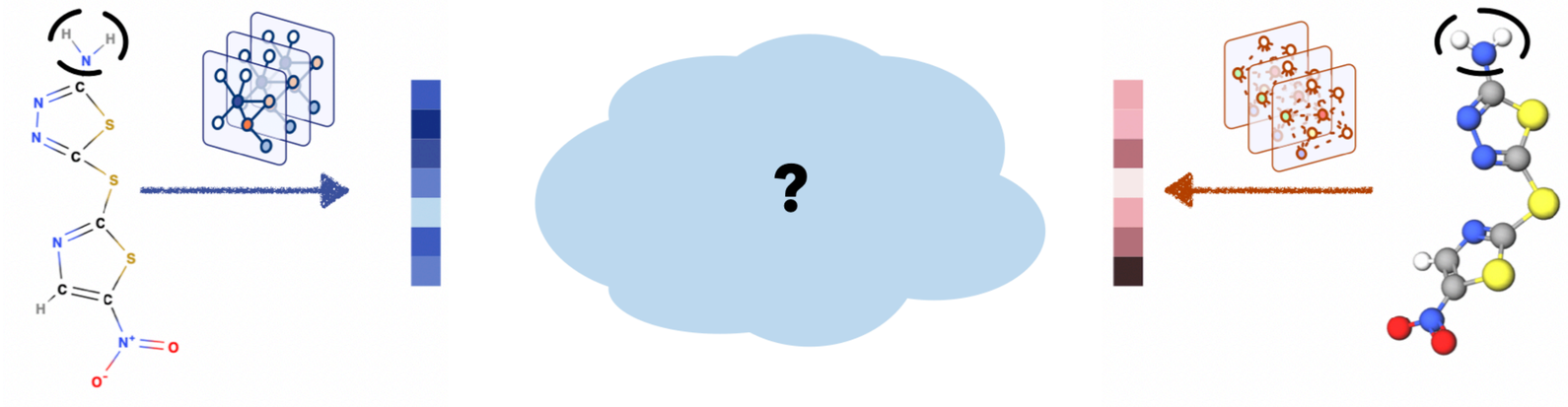
Backgrounds:

- 3D geometric information is useful for molecular property prediction [2, 3], but expensive to obtain via physical experiments or simulation.
- Existing SSL methods on graph are focusing on the 2D topology.

[1] Wu, Zhenqin, et al. "MoleculeNet: a benchmark for molecular machine learning." *Chemical science* 9.2 (2018): 513-530.
[2] Gilmer, Justin, et al. "Neural message passing for quantum chemistry." *International conference on machine learning*. PMLR, 2017.
[3] Liu, Shengchao, Mehmet F. Demirel, and Yingyu Liang. "N-gram graph: Simple unsupervised representation for graphs, with applications to molecules." *Advances in neural information processing systems* 32 (2019).

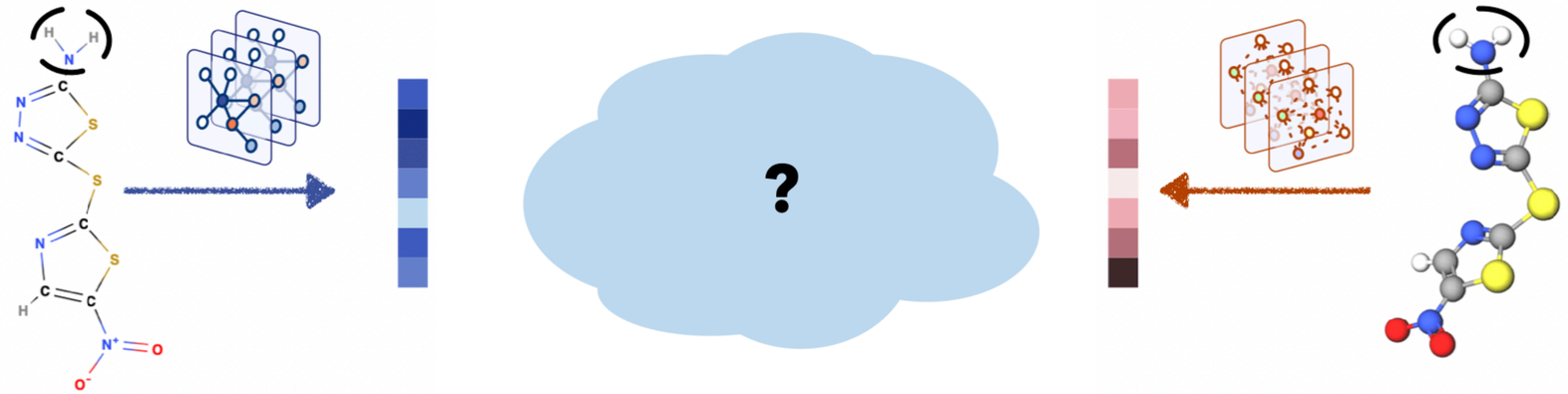
3.2 Motivation

*Q: Suppose we have a **larger/pre-training** dataset with **both 2D and 3D** info, and can we apply **extra 3D info** to help **smaller/downstream** tasks?*



3.2 Motivation

*Q: Suppose we have a **larger/pre-training** dataset with **both 2D and 3D** info, and can we apply **extra 3D info** to help **smaller/downstream** tasks?*



A: We adopt these two views (2D and 3D) and propose **Graph Multi-View Pre-training (GraphMVP)**.

- Pre-training: propose two SSL tasks on both 2D and 3D graph.
- Fine-tuning: downstream tasks with 2D graph only.

3.3 Related Work

SSL Pre-training	View Selection		SSL Category	
	2D Topology	3D Geometry	Generative	Contrastive
EdgePred [1]	✓		✓	
AttrMask [2]	✓		✓	
GPT-GNN [3]	✓		✓	
InfoGraph [4]	✓			✓
ContexPred [2]	✓			✓
GraphLoG [5]	✓			✓
GraphCL [6]	✓			✓
JOAO [7]	✓			✓
GraphMVP (Ours) [9]	✓	✓	✓	✓

[1] Hamilton, William L., Rex Ying, and Jure Leskovec. "Inductive representation learning on large graphs." *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017.
[2] Hu, Weihua, et al. "Strategies for pre-training graph neural networks." *arXiv preprint arXiv:1905.12265* (2019).
[3] Hu, Ziniu, et al. "Gpt-gnn: Generative pre-training of graph neural networks." Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020.
[4] Sun, Fan-Yun, et al. "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization." *arXiv preprint arXiv:1908.01000* (2019).
[5] Xu, Minghao, et al. "Self-supervised Graph-level Representation Learning with Local and Global Structure." *arXiv preprint arXiv:2106.04113* (2021).
[6] You, Yuning, et al. "Graph contrastive learning with augmentations." *Advances in Neural Information Processing Systems* 33 (2020): 5812-5823.
[7] You, Yuning, et al. "Graph Contrastive Learning Automated." *arXiv preprint arXiv:2106.07594* (2021).
[8] Grover, Rong, Yu, et al. "Self-supervised graph transformer on large-scale molecular data." *arXiv preprint arXiv:2007.02835* (2020).
[9] Liu, Shengchao, et al. "Pre-training Molecular Graph Representation with 3D Geometry." *arXiv preprint arXiv:2110.07728* (2021).

3.4 Preliminaries

Notations:

- A: atom (node) attributes.
- E: bond (edge) attributes.
- R: atom (node) positions.

Molecule as 2D topological graph:

- x for a 2D molecular graph.
- h_x for 2D representation, $h_x = \text{2D-GNN}(A, E)$.

Molecule as 3D geometric graph:

- y for a 3D molecular graph.
- h_y for 3D representation, $h_y = \text{3D-GNN}(A, R)$.
- Conformers.

From [1] Axelrod, Simon, and Rafael Gomez-Bombarelli. "GEOM: Energy-annotated molecular conformations for property prediction and molecular generation." *arXiv preprint arXiv:2006.05531* (2020).

CC(C)OC(=O)CCC/C=C\C[C@H]1[C@@H](O)C[C@@H](O)[C@@H]1CC[C@@H](O)CCc1ccccc1

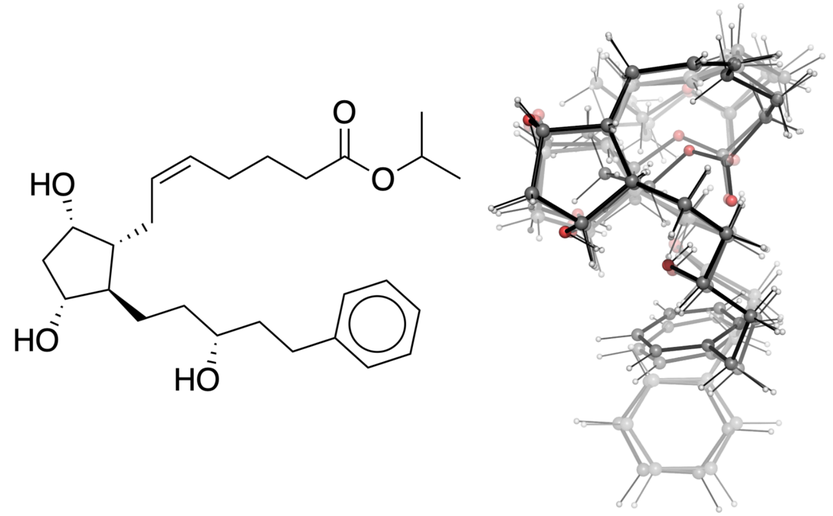


Figure 1. Molecular representations of the latanoprost molecule. *top* SMILES string. *left* Stereochemical formula with edge features, including wedges for in- and out-of-plane bonds, and a double line for *cis* isomerism. *right* Overlay of conformers. Higher transparency corresponds to lower statistical weight.

3.4 Preliminaries

Energy-Based Model (EBM): $p(x) = \frac{\exp(-E(x))}{A}$, where $E(x)$ is the energy function,

and the bottleneck is the intractable partition function $A = \int_x \exp(-E(x))dx$.

Solutions:

- Noise-Contrastive Estimation (NCE) [1, 2]
- Contrastive Divergence
- Score Matching
- …

[1] Liu, Shengchao, et al. "Pre-training Molecular Graph Representation with 3D Geometry." *arXiv preprint arXiv:2110.07728* (2021).

[2] Gutmann, Michael, and Aapo Hyvärinen. "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models." *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010.

3.5 Method: GraphMVP

3.5.1 MI and SSL

3.5.2 Contrastive SSL

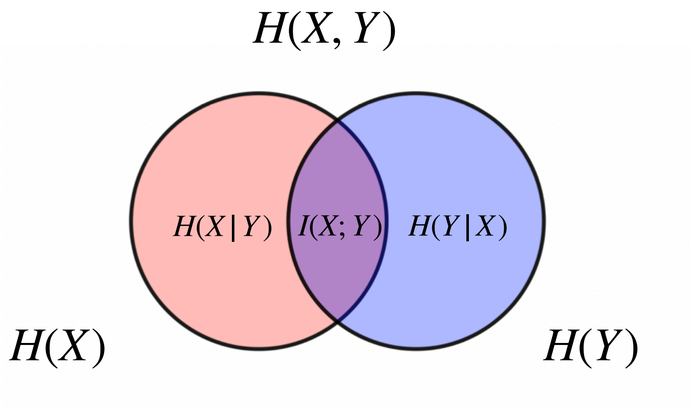
3.5.3 Generative SSL

3.5.4 Multi-task Objective

3.5.1 MI and SSL

Mutual information (MI):

- measures the non-linear dependence between variables.
- the larger MI, the stronger dependence between variables.



Maximizing MI between 2D and 3D views.

- Expect: obtain a more robust 2D representation by sharing more information with its 3D counterparts.

3.5.1 MI and SSL

$$
\begin{aligned}
I(X;Y) &= \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right] \\
&\geq \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{\sqrt{p(x)p(y)}}\right] \\
&= \frac{1}{2}\mathbb{E}_{p(x,y)}\left[\log \frac{(p(x,y))^2}{p(x)p(y)}\right] \\
&= \frac{1}{2}\mathbb{E}_{p(x,y)}\left[\log p(x\,|\,y)\right] + \frac{1}{2}\mathbb{E}_{p(x,y)}\left[\log p(y\,|\,x)\right].
\end{aligned}
$$

How to estimate this lower bound?

3.5.1 MI and SSL

$$
\begin{aligned}
I(X;Y) &= \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right] \\
&\geq \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{\sqrt{p(x)p(y)}}\right] \\
&= \frac{1}{2}\mathbb{E}_{p(x,y)}\left[\log \frac{(p(x,y))^2}{p(x)p(y)}\right] \\
&= \frac{1}{2}\mathbb{E}_{p(x,y)}\Big[\log p(x\,|\,y)\Big] + \frac{1}{2}\mathbb{E}_{p(x,y)}\Big[\log p(y\,|\,x)\Big] \,.
\end{aligned}
$$

How to estimate this lower bound?

GraphMVP proposes 1 contrastive and 1 generative SSL to estimate it, mainly on modeling the conditional log-likelihood term.

3.5.2 Contrastive SSL

Lower bound on MI:

$$I(X;Y) \geq \frac{1}{2}\mathbb{E}_{p(x,y)}[\log p(x\,|\,y) + \log p(y\,|\,x)].$$

If we model the conditional log-likelihood term with energy-based model (EBM):

$$\mathscr{L}_{\text{EBM}} = -\frac{1}{2}\mathbb{E}_{p(x,y)}\left[\log \frac{f_x(x,y)}{A_{x|y}} + \log \frac{f_y(y,x)}{A_{y|x}}\right].$$

3.5.2 Contrastive SSL

Lower bound on MI:

$$I(X;Y) \geq \frac{1}{2}\mathbb{E}_{p(x,y)}[\log p(x\,|\,y) + \log p(y\,|\,x)].$$

If we model the conditional log-likelihood term with energy-based model (EBM):

$$\mathscr{L}_{\text{EBM}} = -\frac{1}{2}\mathbb{E}_{p(x,y)}\left[\log \frac{f_x(x,y)}{A_{x|y}} + \log \frac{f_y(y,x)}{A_{y|x}}\right].$$

Then with NCE, we have the final objective as EBM-NCE:

$$\mathscr{L}_{\text{EBM-NCE}} = -\frac{1}{2}\mathbb{E}_{p_{data}(y)}\left[\mathbb{E}_{p_n(x|y)}[\log(1-\sigma(f_x(x,y)))] + \mathbb{E}_{p_{data}(x|y)}[\log \sigma(f_x(x,y))]\right]$$
$$-\frac{1}{2}\mathbb{E}_{p_{data}(x)}\left[\mathbb{E}_{p_n(y|x)}[\log(1-\sigma(f_y(y,x)))] + \mathbb{E}_{p_{data}(y|x)}[\log \sigma(f_y(y,x))]\right],$$

where p_{data} is the data distribution, p_n is the noise distribution, $f_x(x,y) = f_y(y,x) = \langle h_x, h_y \rangle$.

3.5.2 Contrastive SSL

Lower bound on MI:

$$I(X;Y) \geq \frac{1}{2}\mathbb{E}_{p(x,y)}[\log p(x\,|\,y) + \log p(y\,|\,x)].$$

If we model the conditional log-likelihood term with energy-based model (EBM):

$$\mathscr{L}_{\text{EBM}} = -\frac{1}{2}\mathbb{E}_{p(x,y)}\left[\log \frac{f_x(x,y)}{A_{x|y}} + \log \frac{f_y(y,x)}{A_{y|x}}\right].$$

Then with NCE, we have the final objective as EBM-NCE:

$$\mathscr{L}_{\text{EBM-NCE}} = -\frac{1}{2}\mathbb{E}_{p_{data}(y)}\left[\mathbb{E}_{p_n(x|y)}[\log(1-\sigma(f_x(x,y)))] + \mathbb{E}_{p_{data}(x|y)}[\log\sigma(f_x(x,y))]\right]$$
$$-\frac{1}{2}\mathbb{E}_{p_{data}(x)}\left[\mathbb{E}_{p_n(y|x)}[\log(1-\sigma(f_y(y,x)))] + \mathbb{E}_{p_{data}(y|x)}[\log\sigma(f_y(y,x))]\right],$$

where p_{data} is the data distribution, p_n is the noise distribution, $f_x(x,y) = f_y(y,x) = \langle h_x, h_y \rangle$.

3.5.2 Contrastive SSL

EBM-NCE & Jensen-Shannon Estimation (JSE)

The formulations are similar, while there are 3 main differences:

- Derivation and intuition:
  - JSE: f-divergence, variational estimation, Fenchel duality.
  - EBM-NCE: MI lower bound, EBM, NCE.
- Noise distribution:
  - JSE: MINE [1], empirical distribution for noise distribution.
  - EBM-NCE: recent work [2] extends it with adaptively learnable noise distribution.
- Flexibility:
  - EBM: score matching, contrastive divergence, etc.

[1] Belghazi, Mohamed Ishmael, et al. "Mine: mutual information neural estimation." *arXiv preprint arXiv:1801.04062* (2018).

[2] Arbel, Michael, Liang Zhou, and Arthur Gretton. "Generalized energy based models." *arXiv preprint arXiv:2003.05033* (2020).

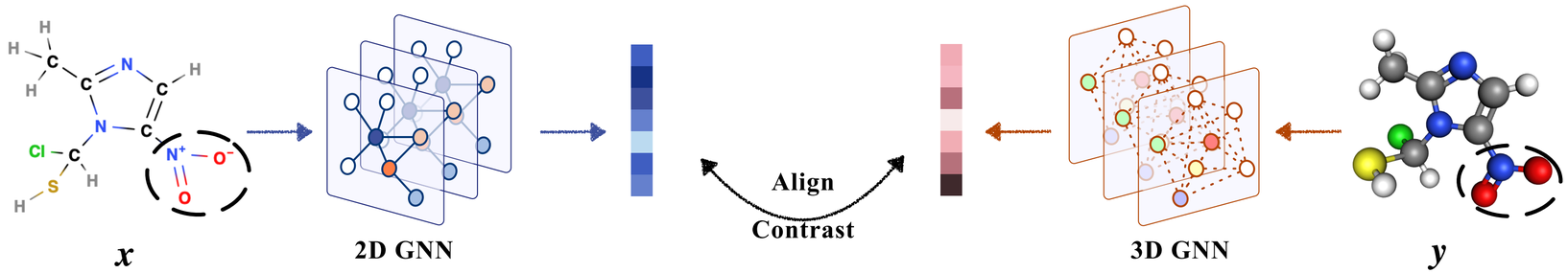
3.5.2 Contrastive SSL

EBM-NCE & InfoNCE

Both EBM-NCE and InfoNCE are aligning the positive pairs and contrasting the negative pairs.

Take either one of them for contrastive SSL, i.e.,

$$\mathscr{L}_{\mathrm{C}} = \mathscr{L}_{\mathrm{InfoNCE}} \text{ or } \mathscr{L}_{\mathrm{C}} = \mathscr{L}_{\mathrm{EBM\text{-}NCE}}.$$



3.5.3 Generative SSL

Lower bound on MI:

$$I(X;Y) \geq \frac{1}{2}\mathbb{E}_{p(x,y)}[\log p(x\,|\,y) + \log p(y\,|\,x)].$$

Variational Molecule Reconstruction

We introduce a variational distribution $z_x = \mu_x + \Sigma_x \odot \epsilon$:

$$\log p(y\,|\,x) = \log \mathbb{E}_{p(z_x)}[p(y\,|\,x, z_x)] \geq \mathbb{E}_{q(z_x|x)}\left[\log p(y\,|\,x, z_x)\right] - KL(q(z_x\,|\,x)\,||\,p(z_x)).$$

3.5.3 Generative SSL

From [1] Axelrod, Simon, and Rafael Gomez-Bombarelli. "GEOM: Energy-annotated molecular conformations for property prediction and molecular generation." *arXiv preprint arXiv:2006.05531* (2020).

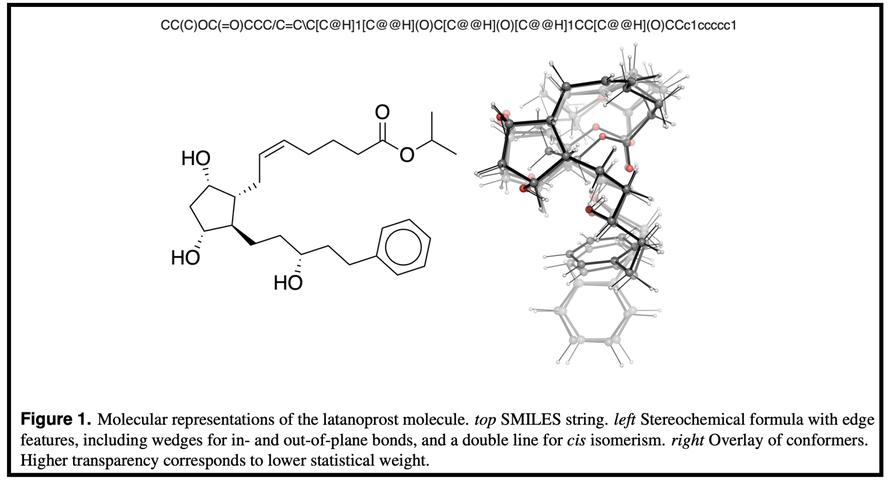


Figure 1. Molecular representations of the latanoprost molecule. *top* SMILES string. *left* Stereochemical formula with edge features, including wedges for in- and out-of-plane bonds, and a double line for *cis* isomerism. *right* Overlay of conformers. Higher transparency corresponds to lower statistical weight.

Lower bound on MI:

$$I(X;Y) \geq \frac{1}{2}\mathbb{E}_{p(x,y)}[\log p(x\,|\,y) + \log p(y\,|\,x)].$$

Variational Molecule Reconstruction

We introduce a variational distribution $z_x = \mu_x + \Sigma_x \odot \epsilon$:

$$\log p(y\,|\,x) = \log \mathbb{E}_{p(z_x)}[p(y\,|\,x, z_x)] \geq \mathbb{E}_{q(z_x|x)}\left[\log p(y\,|\,x, z_x)\right] - KL(q(z_x\,|\,x)\,||\,p(z_x)).$$

Benefits:

- Stochastic mapping between 2D and 3D views.
- An explicit representation for transferring to downstream tasks.

3.5.3 Generative SSL

Lower bound on MI:

$$I(X;Y) \geq \frac{1}{2}\mathbb{E}_{p(x,y)}[\log p(x\,|\,y) + \log p(y\,|\,x)].$$

Variational Molecule Reconstruction

We introduce a variational distribution $z_x = \mu_x + \Sigma_x \odot \epsilon$:

$$\log p(y\,|\,x) = \log \mathbb{E}_{p(z_x)}[p(y\,|\,x, z_x)] \geq \mathbb{E}_{q(z_x|x)}\left[\log p(y\,|\,x, z_x)\right] - KL(q(z_x\,|\,x)\,||\,p(z_x)).$$

Reconstruction

Limitation:

- Reconstruction of structured data. If the target data space is discrete/structured, then the modeling and evaluation on this data space is hard.

3.5.3 Generative SSL

Lower bound on MI:

$$I(X;Y) \geq \frac{1}{2}\mathbb{E}_{p(x,y)}[\log p(x\,|\,y) + \log p(y\,|\,x)].$$

Variational Molecule Reconstruction

We introduce a variational distribution $z_x = \mu_x + \Sigma_x \odot \epsilon$:

$$\log p(y\,|\,x) = \log \mathbb{E}_{p(z_x)}[p(y\,|\,x, z_x)] \geq \mathbb{E}_{q(z_x|x)}\left[\log p(y\,|\,x, z_x)\right] - KL(q(z_x\,|\,x)\,||\,p(z_x)).$$

Reconstruction

Solution:

Variational Representation Reconstruction (VRR)

Let's transfer the reconstruction from **data space** to **representation space**.

3.5.3 Generative SSL

Variational Molecule Reconstruction

$$\log p(y \mid x) = \log \mathbb{E}_{p(z_x)}[p(y \mid x, z_x)] \geq \mathbb{E}_{q(z_x|x)}\left[\log p(y \mid x, z_x)\right] - KL(q(z_x \mid x) \mid\mid p(z_x)).$$

Reconstruction

Variational Representation Reconstruction

Let's transfer the reconstruction from **data space** to **representation space**.

If is y is continuous, we can use Gaussian for the likelihood: $\|y - g_x(z_x)\|^2$, where $g_x(z_x)$ is the decoder.

If is y is discrete and structured, then we propose this surrogate loss: $\|h_y(y) - h_y(g_x(z_x))\|^2$, where h_y is the encoder on y.

By approximation: $\|h_y(y) - q_x(z_x))\|^2$.

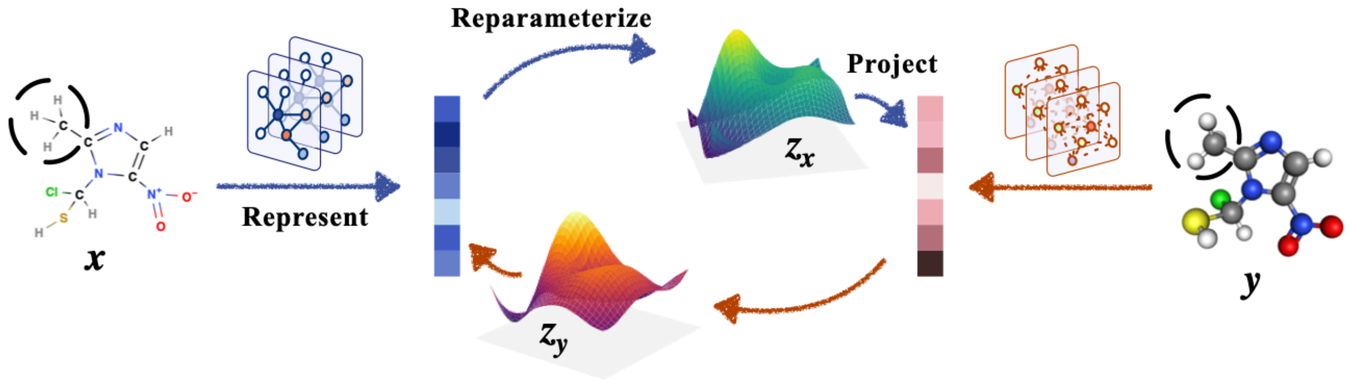
Add stop-gradient: $\|\text{SG}(h_y(y)) - q_x(z_x))\|^2$.

3.5.3 Generative SSL

Final solution (VRR):

$$\mathscr{L}_{\text{G}} = \mathscr{L}_{\text{VRR}} = \frac{1}{2}\left[\mathbb{E}_{q(z_x|x)}\left[\|q_x(z_x) - \text{SG}(h_y)\|^2\right] + \mathbb{E}_{q(z_y|y)}\left[\|q_y(z_y) - \text{SG}(h_x)\|_2^2\right]\right]$$

$$+\frac{\beta}{2} \cdot \left[KL(q(z_x\,|\,x)\,||\,p(z_x)) + KL(q(z_y\,|\,y)\,||\,p(z_y))\right].$$



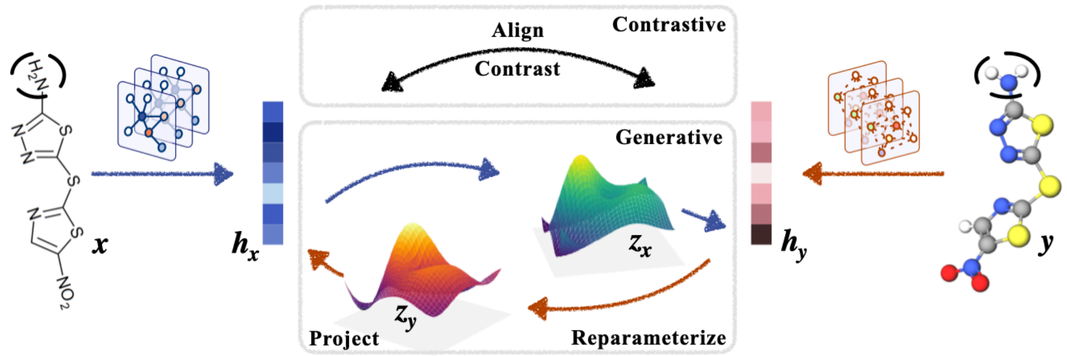
Note: this surrogate loss is exact if h_x/h_y is continuous invertible.

But empirically, we find GNN is good enough.

3.5.4 Multi-task Objective

The objective is weighted sum of the contrastive and generative SSL:

$\mathscr{L}_{\text{GraphMVP}} = \alpha_1 \cdot \mathscr{L}_{\text{C}} + \alpha_2 \cdot \mathscr{L}_{\text{G}}$.



Contrastive and generative SSL are complementary.

- From representation learning:
  - Contrastive SSL is inter-data.
  - Generative SSL is intra-data.
- From distribution learning:
  - Contrastive SSL is learning distribution in a local way: by contrasting negative pairs.
  - Generative SSL is learning distribution in a global way: learning the data density function directly.

3.6 Experiments

Datasets:

- Pre-training
  - GEOM [1], 50k molecules, each with 5 conformers.
- Downstream
  - Molecular Property Prediction:
    - Physiology: Tox21, ToxCast, ClinTox, BBBP, Sider.
    - Physical chemistry: ESOL, Lipophilicity, CEP.
    - Biophysics: MUV, BACE, Hiv, Malaria.
  - Drug-Target Interaction:
    - Davis, KIBA.

Table 8: Summary for the molecule chemical datasets.

Dataset	Task	# Tasks	# Molecules	# Proteins	# Molecule-Protein
BBBP	Classification	1	2,039		
Tox21	Classification	12	7,831		
ToxCast	Classification	617	8,576		
Sider	Classification	27	1,427		
ClinTox	Classification	2	1,478		
MUV	Classification	17	93,087		
HIV	Classification	1	41,127		
Bace	Classification	1	1,513		
Delaney	Regression	1	1,128		
Lipo	Regression	1	4,200		
Malaria	Regression	1	9,999		
CEP	Regression	1	29,978		
Davis	Regression	1	68	379	30,056
KIBA	Regression	1	2,068	229	118,254

Backbone models:

- GIN [2] for 2D GNN.
- SchNet [3] for 3D GNN.

[1] Axelrod, Simon, and Rafael Gomez-Bombarelli. "GEOM: Energy-annotated molecular conformations for property prediction and molecular generation." *arXiv preprint arXiv:2006.05531* (2020).
[2] Xu, Keyulu, et al. "How powerful are graph neural networks?." *arXiv preprint arXiv:1810.00826* (2018).
[3] Schütt, Kristof T., et al. "Schnet–a deep learning architecture for molecules and materials." *The Journal of Chemical Physics* 148.24 (2018): 241722.

3.6 Experiments

Table 1: Results for molecular property prediction tasks. For each downstream task, we report the mean (and standard deviation) ROC-AUC of 3 seeds with scaffold splitting. For GraphMVP, we set $M = 0.15$ and $C = 5$. The best and second best results are marked **bold** and **bold**, respectively.

Pre-training	BBBP	Tox21	ToxCast	Sider	ClinTox	MUV	HIV	Bace	Avg
–	65.4(2.4)	74.9(0.8)	61.6(1.2)	58.0(2.4)	58.8(5.5)	71.0(2.5)	75.3(0.5)	72.6(4.9)	67.21
EdgePred	64.5(3.1)	74.5(0.4)	60.8(0.5)	56.7(0.1)	55.8(6.2)	73.3(1.6)	75.1(0.8)	64.6(4.7)	65.64
AttrMask	70.2(0.5)	74.2(0.8)	62.5(0.4)	60.4(0.6)	68.6(9.6)	73.9(1.3)	74.3(1.3)	77.2(1.4)	70.16
GPT-GNN	64.5(1.1)	**75.3(0.5)**	62.2(0.1)	57.5(4.2)	57.8(3.1)	76.1(2.3)	75.1(0.2)	77.6(0.5)	68.27
InfoGraph	69.2(0.8)	73.0(0.7)	62.0(0.3)	59.2(0.2)	75.1(5.0)	74.0(1.5)	74.5(1.8)	73.9(2.5)	70.10
ContextPred	71.2(0.9)	73.3(0.5)	62.8(0.3)	59.3(1.4)	73.7(4.0)	72.5(2.2)	75.8(1.1)	78.6(1.4)	70.89
GraphLoG	67.8(1.7)	73.0(0.3)	62.2(0.4)	57.4(2.3)	62.0(1.8)	73.1(1.7)	73.4(0.6)	78.8(0.7)	68.47
G-Contextual	70.3(1.6)	75.2(0.3)	62.6(0.3)	58.4(0.6)	59.9(8.2)	72.3(0.9)	75.9(0.9)	79.2(0.3)	69.21
G-Motif	66.4(3.4)	73.2(0.8)	62.6(0.5)	60.6(1.1)	77.8(2.0)	73.3(2.0)	73.8(1.4)	73.4(4.0)	70.14
GraphCL	67.5(3.3)	75.0(0.3)	62.8(0.2)	60.1(1.3)	78.9(4.2)	**77.1(1.0)**	75.0(0.4)	68.7(7.8)	70.64
JOAO	66.0(0.6)	74.4(0.7)	62.7(0.6)	60.7(1.0)	66.3(3.9)	77.0(2.2)	**76.6(0.5)**	72.9(2.0)	69.57
GraphMVP	68.5(0.2)	74.5(0.4)	62.7(0.1)	**62.3(1.6)**	**79.0(2.5)**	75.0(1.4)	74.8(1.4)	76.8(1.1)	71.69
GraphMVP-G	**70.8(0.5)**	**75.9(0.5)**	**63.1(0.2)**	60.2(1.1)	**79.1(2.8)**	**77.7(0.6)**	76.0(0.1)	**79.3(1.5)**	**72.76**
GraphMVP-C	**72.4(1.6)**	74.4(0.2)	**63.1(0.4)**	**63.9(1.2)**	77.5(4.2)	75.0(1.0)	**77.0(1.2)**	**81.2(0.9)**	**73.07**

Table 5: Results for four molecular property prediction tasks (regression) and two DTA tasks (regression). We report the mean RMSE of 3 seeds with scaffold splitting for molecular property downstream tasks, and mean MSE for 3 seeds with random splitting on DTA tasks. For GraphMVP, we set $M = 0.15$ and $C = 5$. The best performance for each task is marked in **bold**. We omit the std here since they are very small and indistinguishable. For complete results, please check Appendix G.4.

	Molecular Property Prediction					Drug-Target Affinity		
Pre-training	ESOL	Lipo	Malaria	CEP	Avg	Davis	KIBA	Avg
–	1.178	0.744	1.127	1.254	1.0756	0.286	0.206	0.2459
AM	1.112	0.730	1.119	1.256	1.0542	0.291	0.203	0.2476
CP	1.196	0.702	1.101	1.243	1.0606	0.279	0.198	0.2382
JOAO	1.120	0.708	1.145	1.293	1.0663	0.281	0.196	0.2387
GraphMVP	1.091	0.718	1.114	1.236	1.0397	0.280	0.178	0.2286
GraphMVP-G	1.064	0.691	1.106	**1.228**	1.0221	**0.274**	0.175	0.2248
GraphMVP-C	**1.029**	**0.681**	**1.097**	1.244	**1.0128**	0.276	**0.168**	**0.2223**

4. MTL for Molecular Property Prediction

4.1 Problem Definition

4.2 Related Work

4.3 Preliminaries

4.4 Dataset with Explicit Task Relation

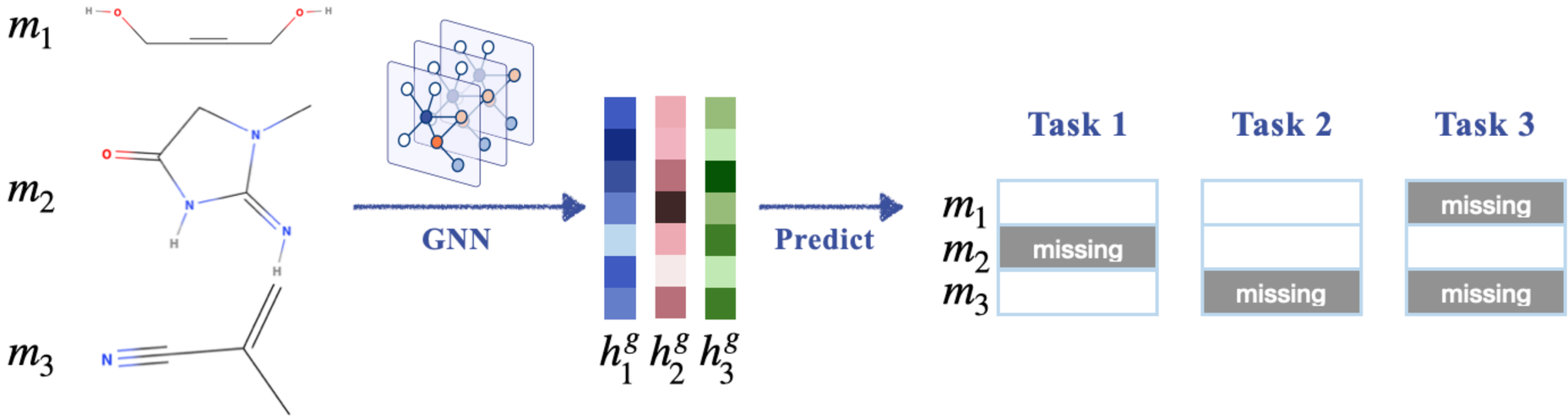
4.5 Method: SGNN-EBM

4.6 Experiments

AI for Science: Mind the Gaps, NeurIPS 2021 Workshop; In Submission to AISTATS'22
Shengchao Liu, *Meng Qu, Zuobai Zhang, Huiyu Cai, Jian Tang*

4.1 Problem Definition

Molecule -> Shared Representation -> Property Prediction on Multiple Tasks



4.1 Problem Definition

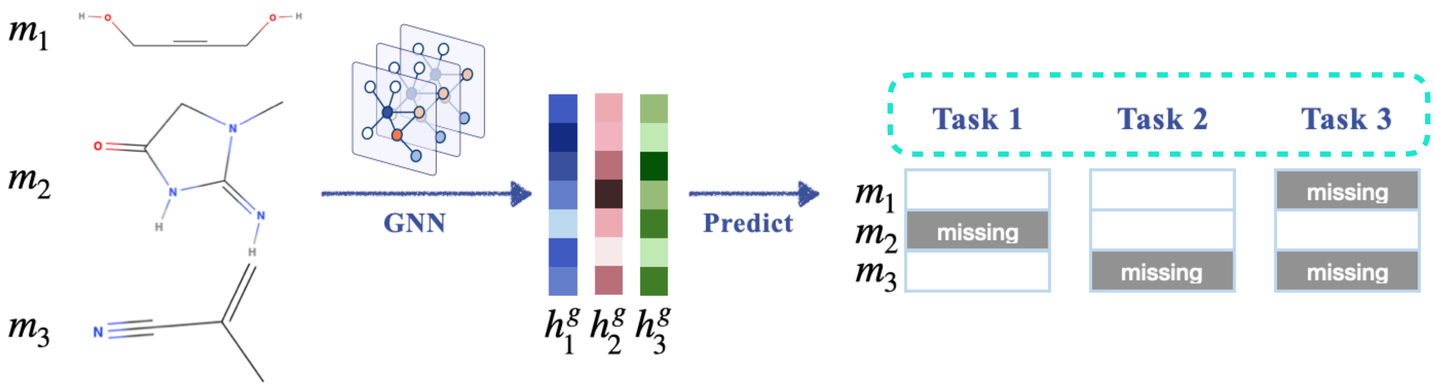
Molecule -> Shared Representation -> Property Prediction on Multiple Tasks

- Useful tool for low-data & missing labels.

- The domain knowledge is rich, can we take better advantage of them?
  - What format of domain knowledge we can utilize?
  - With the specific format of domain knowledge, how to incorporate them for problem solving?

4.1 Problem Definition

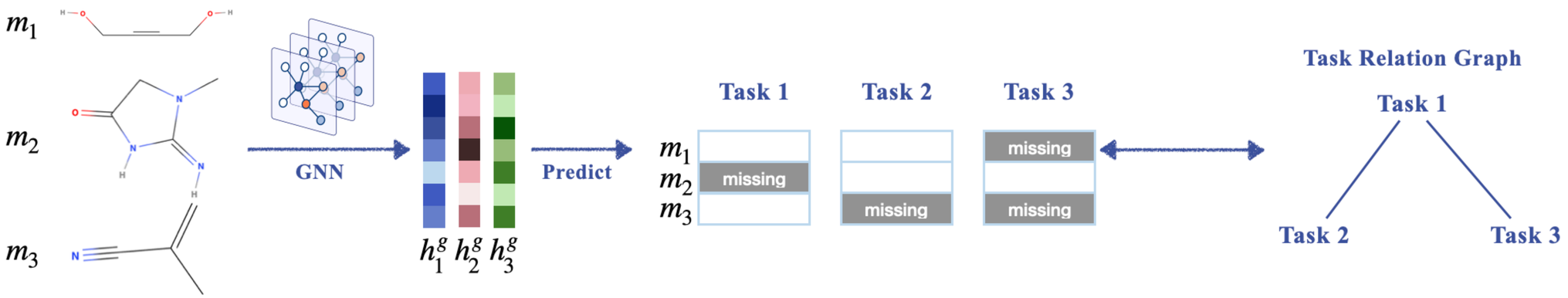
Q: What format of domain knowledge we can utilize?



4.1 Problem Definition

Q: What format of domain knowledge we can utilize?

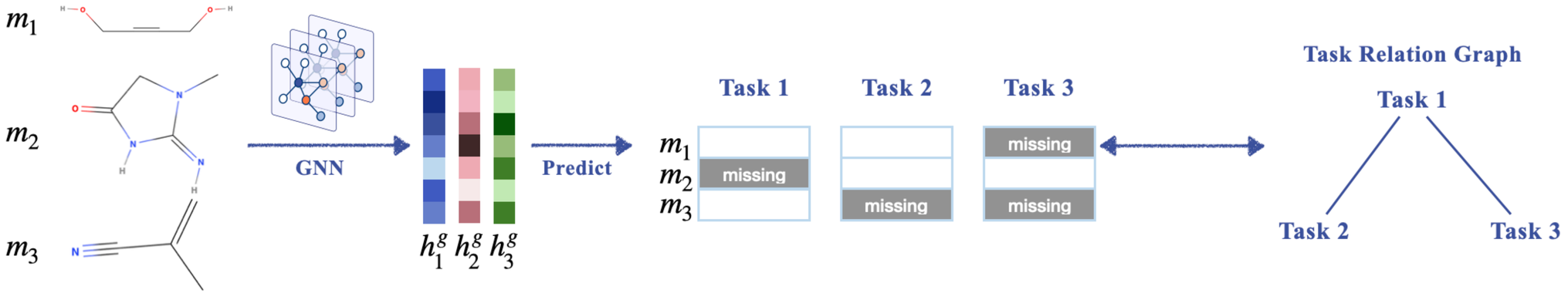
A: We can extract task relation graph from domain.



4.1 Problem Definition

Q: What format of domain knowledge we can utilize?

A: We can extract task relation graph from domain.

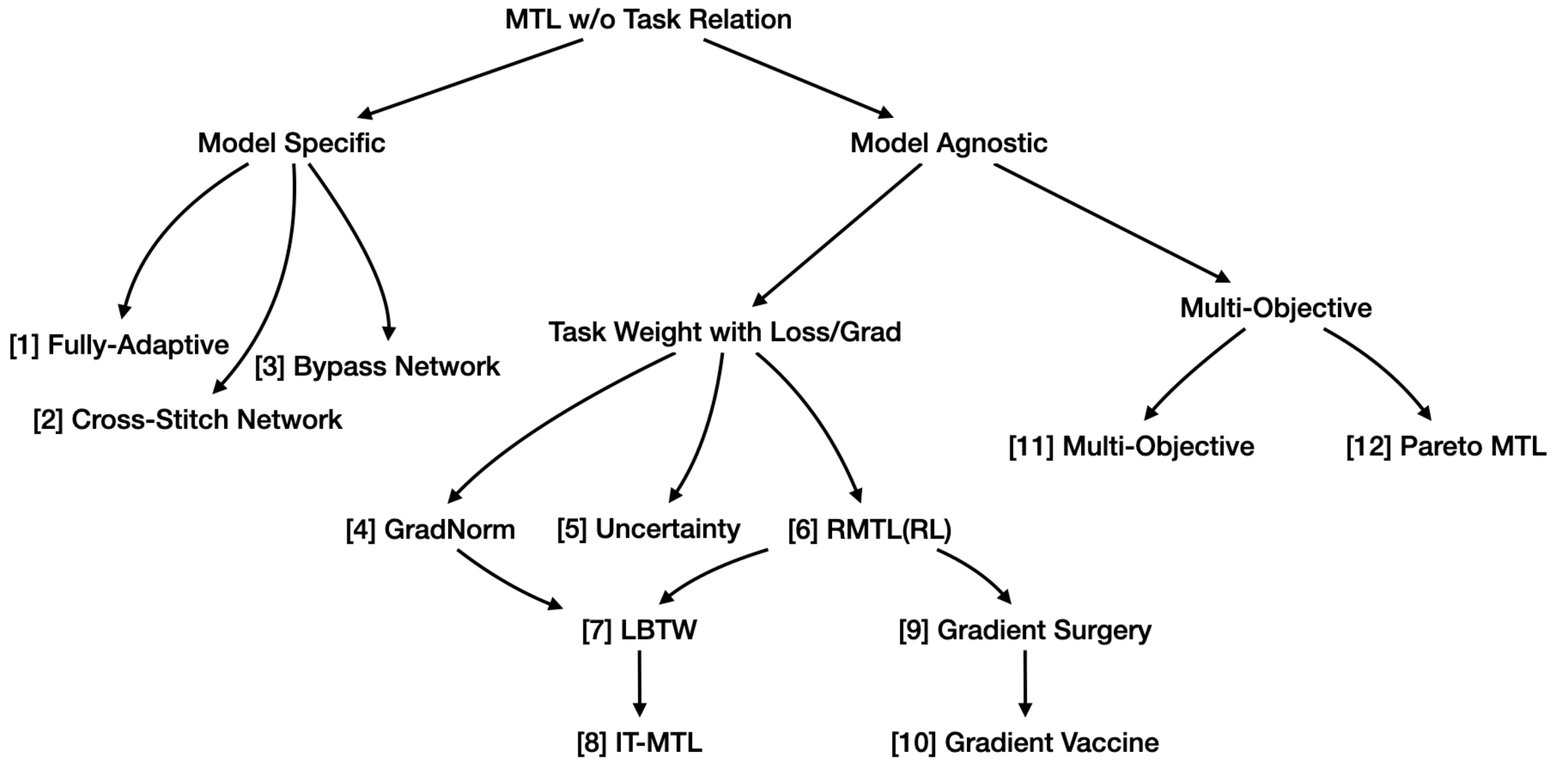


Q: With the specific format of domain knowledge, how to incorporate them for problem solving?

A: We solve this from two directions in modeling the task relation graph:

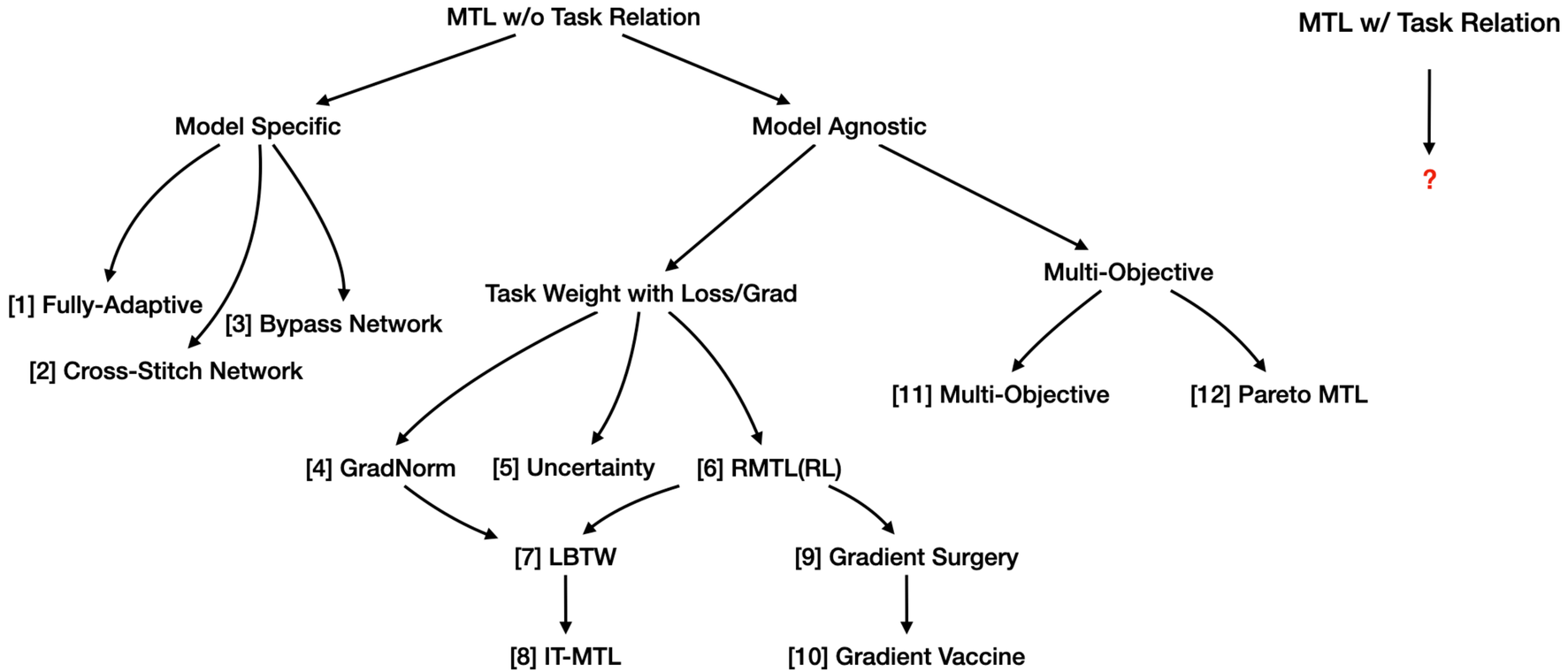
- Direction 1: modeling in the **latent space.**
- Direction 2: modeling in the **output space.**

4.2 Related Work



[1] Lu, Yongxi, et al. "Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
[2] Misra, Ishan, et al. "Cross-stitch networks for multi-task learning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[3] Ramsundar, Bharath, et al. "Is multitask deep learning practical for pharma?." *Journal of chemical information and modeling* 57.8 (2017): 2068-2076.
[4] Chen, Zhao, et al. "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks." *International Conference on Machine Learning*. PMLR, 2018.
[5] Kendall, Alex, Yarin Gal, and Roberto Cipolla. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
[6] Liu, Shengchao. Exploration on deep drug discovery: Representation and learning. 2018.
[7] Liu, Shengchao, Yingyu Liang, and Anthony Gitter. "Loss-balanced task weighting to reduce negative transfer in multi-task learning." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 2019.
[8] Wu, Sen, Hongyang R. Zhang, and Christopher Ré. "Understanding and improving information transfer in multi-task learning." *arXiv preprint arXiv:2005.00944* (2020).
[9] Yu, Tianhe, et al. "Gradient surgery for multi-task learning." *arXiv preprint arXiv:2001.06782* (2020).
[10] Wang, Zirui, et al. "Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models." *arXiv preprint arXiv:2010.05874* (2020).
[11] Sener, Ozan, and Vladlen Koltun. "Multi-task learning as multi-objective optimization." *arXiv preprint arXiv:1810.04650* (2018).
[12] Lin, Xi, et al. "Pareto multi-task learning." Advances in neural information processing systems 32 (2019): 12060-12070.

4.2 Related Work



[1] Lu, Yongxi, et al. "Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
[2] Misra, Ishan, et al. "Cross-stitch networks for multi-task learning." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[3] Ramsundar, Bharath, et al. "Is multitask deep learning practical for pharma?." *Journal of chemical information and modeling* 57.8 (2017): 2068-2076.
[4] Chen, Zhao, et al. "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks." *International Conference on Machine Learning*. PMLR, 2018.
[5] Kendall, Alex, Yarin Gal, and Roberto Cipolla. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.
[6] Liu, Shengchao. Exploration on deep drug discovery: Representation and learning. 2018.
[7] Liu, Shengchao, Yingyu Liang, and Anthony Gitter. "Loss-balanced task weighting to reduce negative transfer in multi-task learning." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 2019.
[8] Wu, Sen, Hongyang R. Zhang, and Christopher Ré. "Understanding and improving information transfer in multi-task learning." *arXiv preprint arXiv:2005.00944* (2020).
[9] Yu, Tianhe, et al. "Gradient surgery for multi-task learning." *arXiv preprint arXiv:2001.06782* (2020).
[10] Wang, Zirui, et al. "Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models." *arXiv preprint arXiv:2010.05874* (2020).
[11] Sener, Ozan, and Vladlen Koltun. "Multi-task learning as multi-objective optimization." *arXiv preprint arXiv:1810.04650* (2018).
[12] Lin, Xi, et al. "Pareto multi-task learning." Advances in neural information processing systems 32 (2019): 12060-12070.

4.3 Preliminaries

- molecule: $x = (V, E)$, V is the node attributes, E is the edge attributes.
- T tasks with C-class labels: $y = \{y_0, y_1, \ldots, y_{T-1}\}$, where we focus on $C = 2$.
- Graph Neural Network (GNN): GCN[1], GIN[2].
- Energy-Based Model (EBM):

$$p_\phi(y \mid x) = \frac{\exp(-E_\phi(x, y))}{A}.$$

[1] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." *arXiv preprint arXiv:1609.02907* (2016).
[2] Xu, Keyulu, et al. "How powerful are graph neural networks?." *arXiv preprint arXiv:1810.00826* (2018).

4.4 Dataset with Explicit Task Relation

456k molecules and 1k tasks from ChEMBL: https://www.ebi.ac.uk/chembl/

Task (protein) reference to STRING: https://string-db.org/

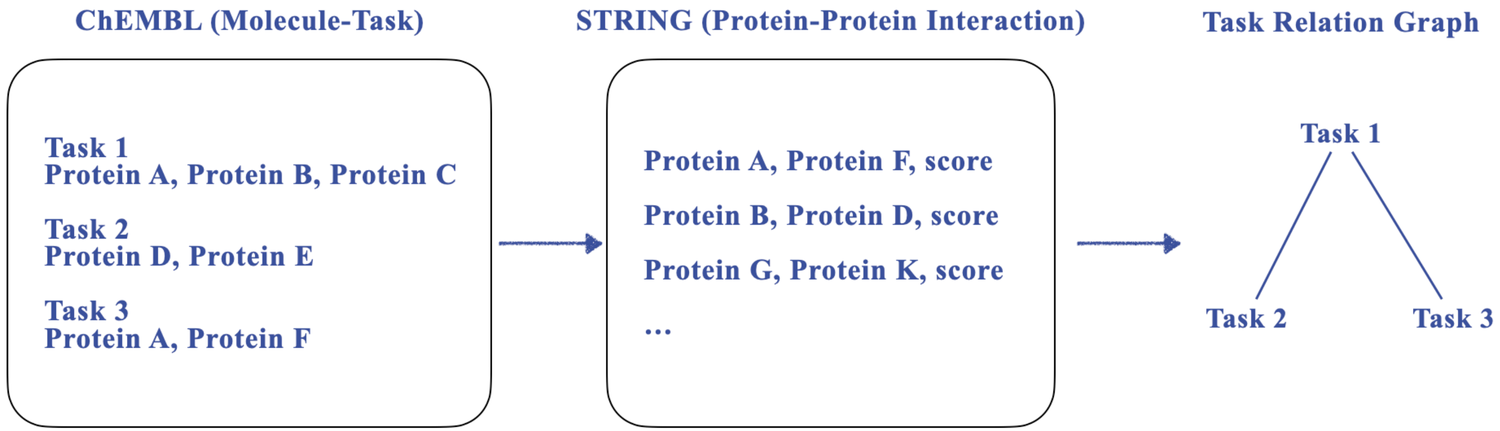


Table 1: Statistics about 3 benchmark datasets with explicit task relation, filtered by 3 thresholds. Threshold means the number of non-missing labels for each molecule/task.

Threshold	# Molecules	# Tasks	Sparsity
10	13,004	382	5.76%
50	932	152	66.70%
100	518	132	92.87%

4.5 Method: SGNN-EBM

4.5.1 Input Embedding

4.5.2 Structured Latent Space Modeling: State Graph Neural Network (SGNN)

4.5.3 Structured Output Space Modeling: Energy-Based Model (EBM)

4.5.4 SGNN-EBM

4.5.1 Input Embedding

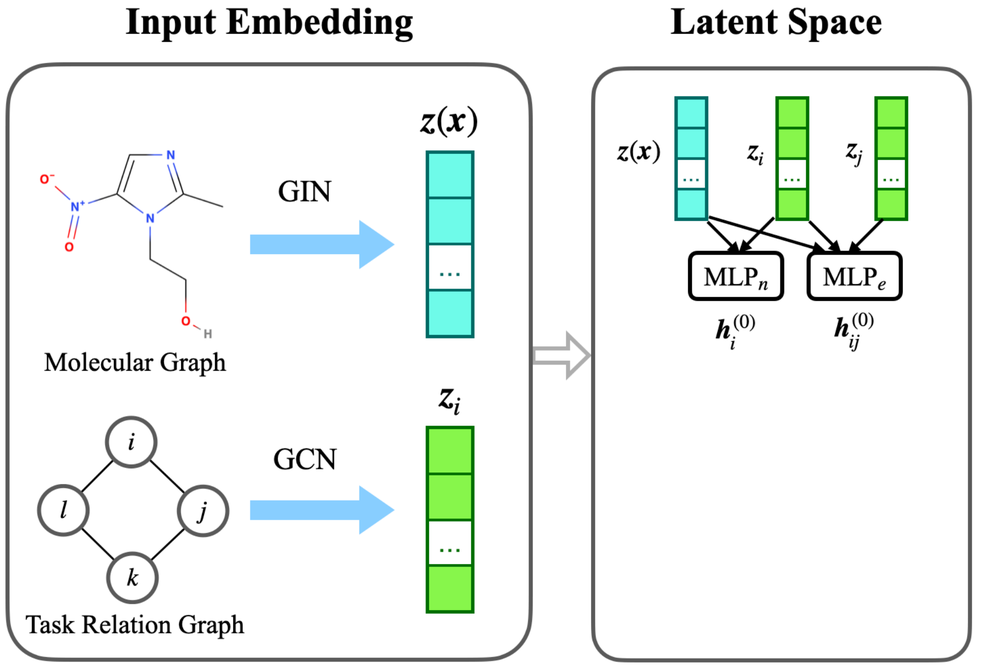
Molecule Embedding:

$z(x) = \text{GIN}(V, E)$, where $z(x) \in \mathbb{R}^{d_m}$.

Task Embedding:

$[z_0, z_1, \ldots, z_{T-1}] = \text{GCN}(\text{task relation graph})$,

where $z_i \in \mathbb{R}^{d_t}$.



Concat these two embeddings as inputs into state GNN (SGNN):

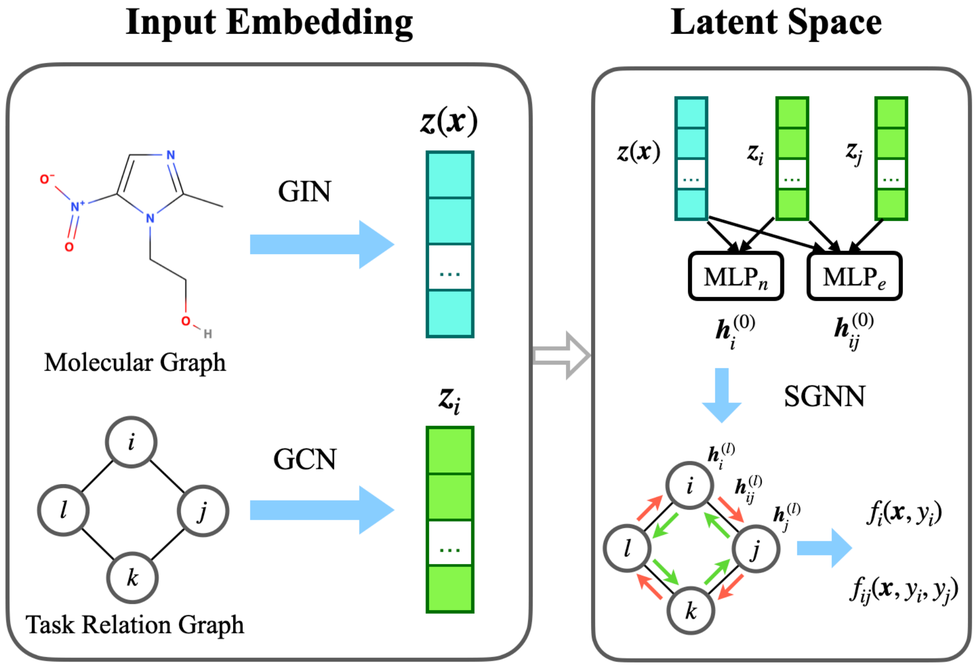
$h_i^{(0)}(x) = \text{MLP}_n^{(0)}(z(x) \oplus z^{(i)})$

$h_{ij}^{(0)}(x) = \text{MLP}_e^{(0)}(z(x) \oplus z^{(i)} \oplus z^{(j)})$,

where $\text{MLP}_n^{(0)} : \mathbb{R}^{d_m+d_t} \to \mathbb{R}^{C\times d}$, $\text{MLP}_e^{(0)} : \mathbb{R}^{d_m+2d_t} \to \mathbb{R}^{C\times C\times d}$.

4.5.2 Structured Latent Space Modeling: SGNN

State Graph Neural Network (SGNN)



- State on the input layer:

$$h_i^{(0)}(x, y_i) = h_i^{(0)}(x)[y_i]$$

$$h_{ij}^{(0)}(x, y_i, y_j) = h_{ij}^{(0)}(x)[y_i, y_j] .$$

- Message passing:

$$h_i^{(l+1)}(x, y_i) = \text{MPNN}_n^{(l+1)}\left(h_i^{(l)}(x, y_i), \left\{ h_{ij}^{(l)}(x, y_i, y_j) \mid \forall j, y_j \right\} \right)$$

$$h_{ij}^{(l+1)}(x, y_i, y_j) = \text{MPNN}_e^{(l+1)}\left(h_i^{(l)}(x, y_i), h_j^{(l)}(x, y_j), h_{ij}^{(l)}(x, y_i, y_j) \right) .$$

4.5.3 Structured Output Space Modeling: EBM

Energy-Based Model (EBM)

- Energy function

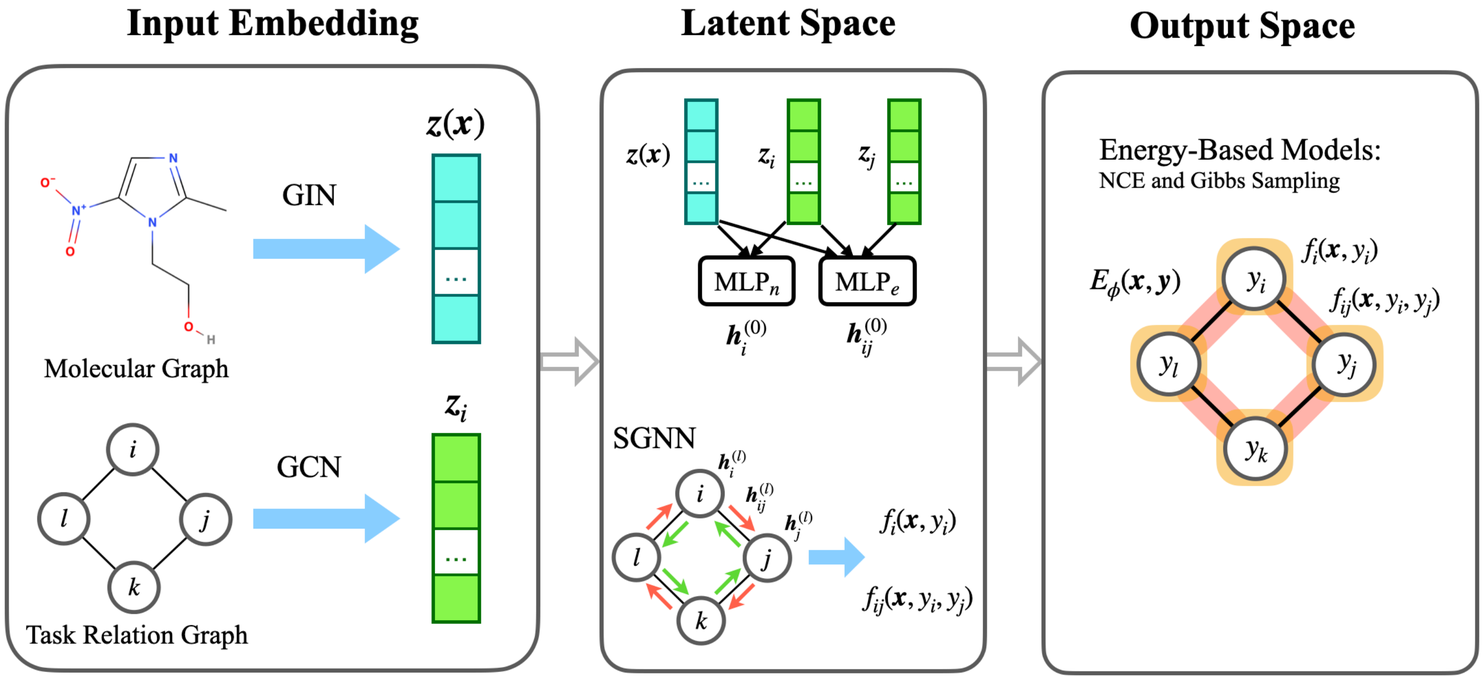
$$E_\phi(x, y) = -\sum_{i=0}^{T-1} f_i(x, y_i) - \lambda \sum_{\langle i,j\rangle \in \mathcal{G}} f_{ij}(x, y_i, y_j)\,.$$

- EBM likelihood

$$p_\phi(y \,|\, x) = \frac{\exp\Big(\sum_i f_i(x, y_i) + \sum_{ij} f_{ij}(x, y_i, y_j)\Big)}{A}\,.$$

4.5.4 SGNN-EBM

- **SGNN** for modeling in the latent space.
- **EBM** for modeling in the output space.
- The final model is called **SGNN-EBM**.



4.5.4 SGNN-EBM

- Learning with EBM-NCE:

$$\mathscr{L}_{NCE} = \mathbb{E}_{y\sim p_n} \log \frac{1}{1+\exp(-E_\phi(x,y))} + \mathbb{E}_{y\sim p_{data}} \log \frac{1}{1+\exp(E_\phi(x,y))}.$$

- Inference with Gibbs sampling:

$$p_\phi(y_i \,|\, y_{-i}, x) = \frac{\exp\left(f_i(x,y_i) + \sum_{\langle i,j\rangle\in\mathscr{G}} f_{ij}(x,y_i,y_j)\right)}{\sum_{y_i=0}^{C-1} \exp\left(f_i(x,y_i) + \sum_{\langle i,j\rangle\in\mathscr{G}} f_{ij}(x,y_i,y_j)\right)}.$$

4.6 Experiments

Empirical results on one dataset with three thresholds.

Table 1: Statistics about 3 benchmark datasets with explicit task relation, filtered by 3 thresholds. Threshold means the number of non-missing labels for each molecule/task.

Threshold	# Molecules	# Tasks	Sparsity
10	13,004	382	5.76%
50	932	152	66.70%
100	518	132	92.87%

Table 2: Main MTL results. All datasets are split into 8-1-1 for train, valid, and test respectively. For each method, we run 5 seeds and report the mean and standard deviation. The best performance is **highlighted**.

Method	p_n	ChEMBL 10	ChEMBL 50	ChEMBL 100
STL	–	71.67 ± 0.64	73.57 ± 1.20	70.81 ± 1.28
MTL	–	74.83 ± 0.61	79.37 ± 1.76	77.78 ± 1.59
UW	–	72.49 ± 0.53	79.68 ± 0.98	78.71 ± 1.93
GradNorm	–	75.17 ± 0.77	79.46 ± 1.27	78.75 ± 1.60
DWA	–	72.45 ± 1.31	79.35 ± 0.68	78.21 ± 2.31
LBTW	–	75.21 ± 0.49	79.52 ± 0.56	79.07 ± 0.99
SGNN	–	77.72 ± 0.66	79.69 ± 1.07	80.19 ± 2.01
SGNN-EBM	SGNN (Fixed)	78.04 ± 0.73	80.34 ± 1.08	80.48 ± 1.93
SGNN-EBM	SGNN (Adaptive)	**78.10 ± 0.71**	**80.78 ± 0.85**	**81.13 ± 2.04**

5. Conclusions & Future Directions

- About SSL on graph:
  - We show that 3D information can help 2D representation. Can we show that 2D information can help 3D representation? E.g., take downstream with 3D only.
  - EBM-NCE connects EBM and SSL, can we try other solutions to EBM?
  - Generative SSL (Variational Representation Reconstruction, VRR) contains the non-contrastive SSL (e.g., BYOL, SimSiam).
    - *Q: If BYOL/SimSiam can provide a robust representation, does this mean other generative SSL can also reach comparative performance?*

5. Conclusions & Future Directions

- About SSL on graph:
  - We show that 3D information can help 2D representation. Can we show that 2D information can help 3D representation? E.g., take downstream with 3D only.
  - EBM-NCE connects EBM and SSL, can we try other solutions to EBM?
  - Generative SSL (Variational Representation Reconstruction, VRR) contains the non-contrastive SSL (e.g., BYOL, SimSiam).
    - *Q: If BYOL/SimSiam can provide a robust representation, does this mean other generative SSL can also reach comparative performance?*
    - A: Yes! This work [1] provides the empirical evidence.

[1] He, Kaiming, et al. "Masked autoencoders are scalable vision learners." *arXiv preprint arXiv:2111.06377* (2021).

5. Conclusions & Future Directions

- About MTL on graph:
  - Can we extend this to different scientific applications?
  - Can we learn such task relation graph?
  - The extracted task relation is noisy. Can we use the learned task relation to help rectify it?

5. Conclusions & Future Directions

- About MTL on graph:
  - Can we extend this to different scientific applications?
  - Can we learn such task relation graph?
  - The extracted task relation is noisy. Can we use the learned task relation to help rectify it?

- More generally:
  - Combining SSL and MTL. Now we explore two directions separately. In the future, we can combine these two directions into a unified pipeline.
  - What other formats of domain knowledge can we utilize? And how to incorporate them appropriately with the AI methods?

Thank you!

I would like to thank all the supports and discussions from Prof. Jian Tang and the Deep Graph Learning Team in MILA: Meng Qu, Zuobai Zhang, Huiyu Cai, Andreea Deac, Vikas Verma, Zhaocheng Zhu, Chence Shi, Minghao Xu, Minkai Xu.

Besides, I would also like to thank all the collaborators during my past and ongoing research projects: (alphabetical order)

Prof. Dimitris Achlioptas, Prof. Anthony Gitter, Prof. Hongyu Guo, Prof. Yingyu Liang, Prof. Jian-Yun Nie, Prof. Dimitris Papailiopoulos.

Moayad Alnammi, Danica Cao, Mehmet F. Demirel, Yuanqi Du, Spencer S. Ericksen, Tianfan Fu, Siddhant Garg, Sai Krishna Gottipati, Yunhe Li, Weiyang Liu, Qi Liu, Pierre-Andre Noel, Dinming Wang, Hanchen Wang, Scott Wildman, David Vazquez, Yutao Zhu.

Q & A