Pre-training Molecular Graph Representation with 3D Geometry — Rethinking Self-Supervised Learning on Structured Data

Shengchao Liu[1,2], Hanchen Wang[3], Weiyang Liu[3,4], Joan Lasenby[3], Hongyu Guo[5], Jian Tang[1,6,7]

[1]Mila, [2]Université de Montréal, [3]University of Cambridge, [4]MPI for Intelligent Systems, Tübingen, [5]National Research Council Canada, [6]HEC Montréal, [7]CIFAR AI Chair

Self-Supervised Learning & Molecular Property Prediction

For molecular property prediction:

	Accessibility	Information
2D Topology	Easy	Low
3D Geometry	Hard	High

Q: Can we find a smarter way to utilize 3D information to help augment the 2D representation?
A: Yes, and we propose Graph Multi-View Pre-training (GraphMVP).

- It uses **both** 3D and 2D in SSL **pre-training**.
- It uses **only** 2D info for downstream **fine-tuning**.

Existing Graph SSL

- Existing SSL on graph.
  - Node-level
  - Context-level
  - Graph-level

SSL Pre-training	Graph View		SSL Category	
	2D Topology	3D Geometry	Generative	Contrastive
EdgePred [31]	✓		✓	
AttrMask [38]	✓		✓	
GPT-GNN [39]	✓		✓	
InfoGraph [71, 79]	✓			✓
ContexPred [38]	✓			✓
GraphLoG [88]	✓			✓
GraphCL [91]	✓			✓
JOAO [90]	✓			✓
GraphMVP (Ours)	✓	✓	✓	✓

Method: GraphMVP

- **Mutual Information (MI) and Self-Supervised Learning (SSL)**

$$I(X;Y) \geq \mathbb{E}_{p(x,y)}\left[\log \frac{p(x,y)}{\sqrt{p(x)p(y)}}\right]$$

$$= \frac{1}{2}\mathbb{E}_{p(x,y)}\Big[\log p(x\,|\,y)\Big] + \frac{1}{2}\mathbb{E}_{p(x,y)}\Big[\log p(y\,|\,x)\Big]$$

- **Contrastive SSL**
  - EBM-NCE

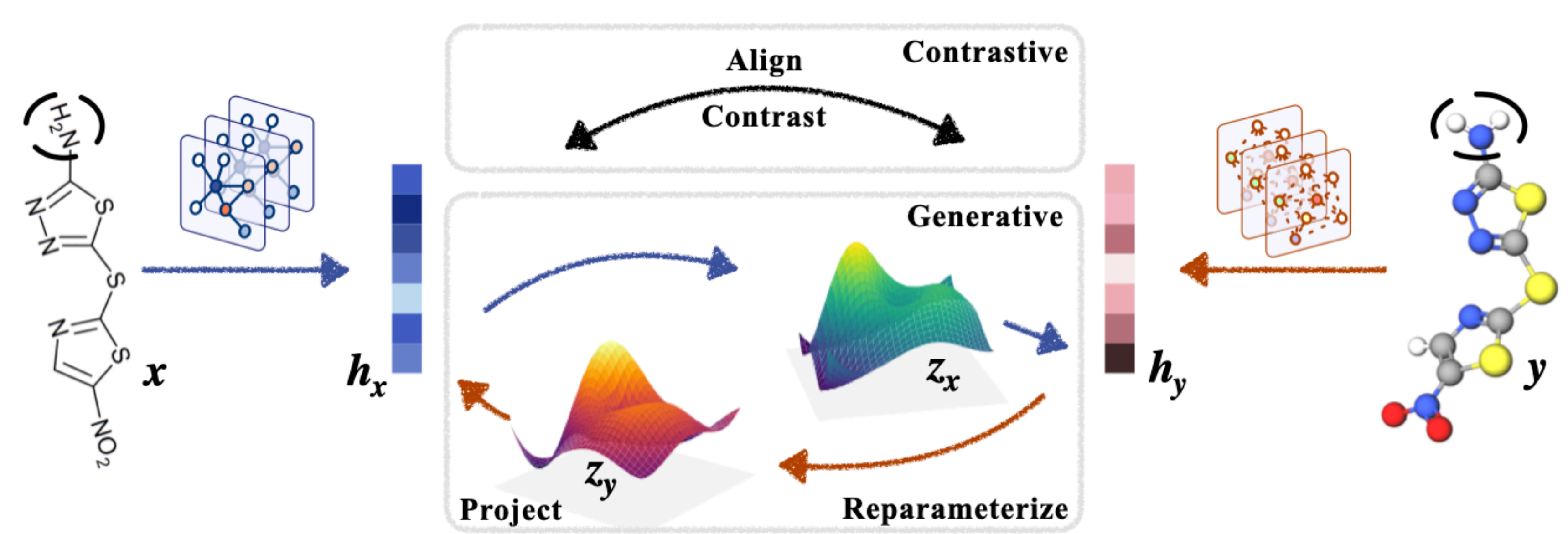
$$\mathscr{L}_{\text{EBM-NCE}} = -\frac{1}{2}\mathbb{E}_{p_{\text{data}}(y)}\Big[\mathbb{E}_{p_n(x|y)}[\log\big(1-\sigma(f_x(x,y))\big)] + \mathbb{E}_{p_{\text{data}}(x|y)}[\log\sigma(f_x(x,y))]\Big]$$
$$-\frac{1}{2}\mathbb{E}_{p_{\text{data}}(x)}\Big[\mathbb{E}_{p_n(y|x)}[\log\big(1-\sigma(f_y(y,x))\big)]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

- **Generative SSL**
  - Which generative model? VAE-like.
  - Continuous & Structured data space
    - Variational Representation Reconstruction (VRR)

$$\mathscr{L}_{\text{G}} = \mathscr{L}_{\text{VRR}} = \frac{1}{2}\Big[\mathbb{E}_{q(z_x|x)}\big[\|q_x(z_x) - \text{SG}(h_y)\|^2\big] + \mathbb{E}_{q(z_y|y)}\big[\|q_y(z_y) - \text{SG}(h_x)\|_2^2\big]\Big]$$
$$+\frac{\beta}{2}\cdot\Big[KL(q(z_x\,|\,x)\,||\,p(z_x)) + KL(q(z_y\,|\,y)\,||\,p(z_y))\Big]$$

    - This is SimSiam with randomness!

- **Multi-task Objectives**
  - Contrastive SSL and Generative SSL are complementary
    - Inter-data and intra-data
    - Local and global
  - Objective: $\mathscr{L}_{\text{GraphMVP}} = \alpha_1 \cdot \mathscr{L}_{\text{C}} + \alpha_2 \cdot \mathscr{L}_{\text{G}}$



Experiments

- Backbone models: GIN for 2D, SchNet for 3D
- Pre-training dataset: GEOM
- Fine-tuning datasets: 8 binary classification datasets

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
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**

- Ablation study on objective function: each individual contrastive and generative SSL.

Table 3: Ablation on the objective function.

GraphMVP Loss	Contrastive	Generative	Avg
Random			67.21
InfoNCE only	✓		68.85
EBM-NCE only	✓		70.15
VRR only		✓	69.29
RR only		✓	68.89
InfoNCE + VRR	✓	✓	70.67
EBM-NCE + VRR	✓	✓	71.69
InfoNCE + RR	✓	✓	70.60
EBM-NCE + RR	✓	✓	70.94

Findings and Conclusions

- Problem novelty:
  - A novel research direction to utilize 3D representation to augment 2D representation, especially for structured data.
- Technical novelty:
  - EBM-NCE: connects energy-based model (EBM) with SSL.
  - VRR: proposes a novel generative SSL; provides another viewpoint for non-contrastive SSL.
  - Contrastive SSL and Generative SSL are complementary.

Codes will be available at https://github.com/chao1224/GraphMVP

Email: liusheng@mila.quebec, arXiv: https://arxiv.org/abs/2110.07728