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

ICLR 2022

Shengchao Liu, *Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, Jian Tang*

Pipeline

1 Motivation & Problem Definition

2 Related Work

3 Preliminaries

4 Method: GraphMVP

5 Experiments

6 Future Directions: SSL on Structured Data

1 Motivation & Problem Definition

Ultimate goal:

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

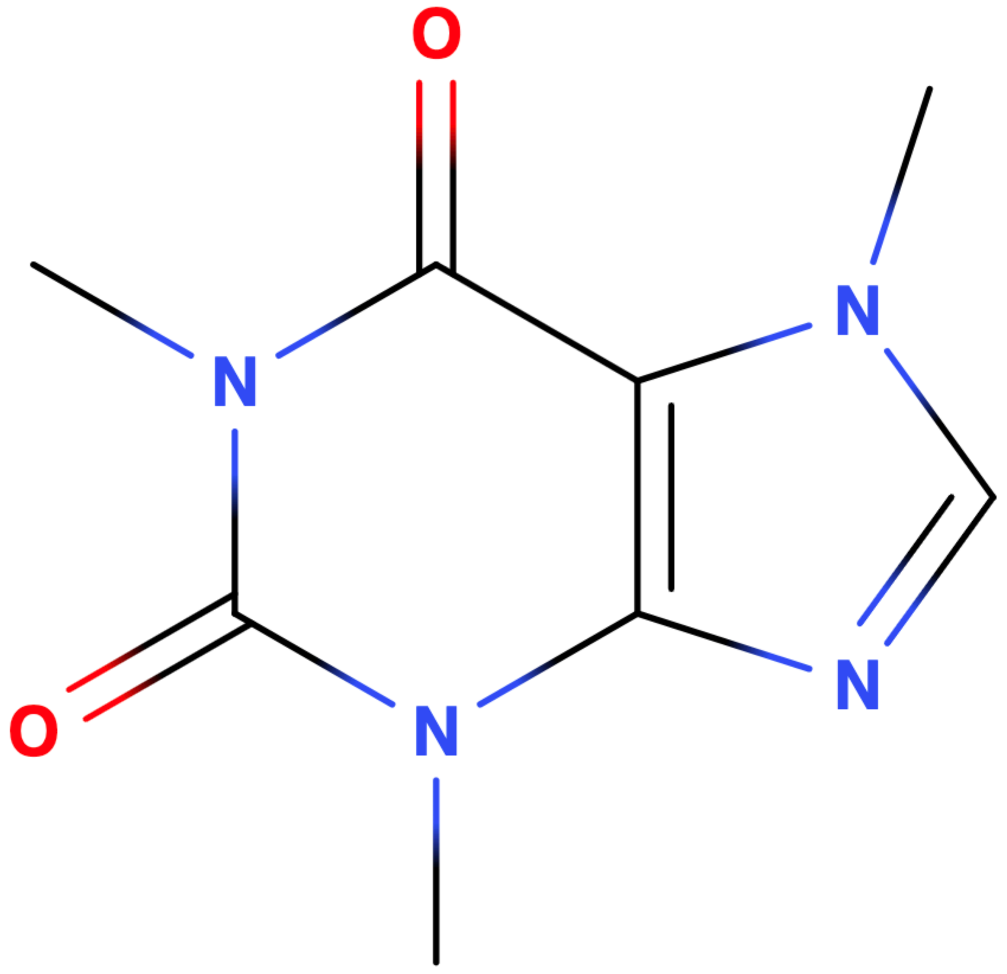
[1] Wu, Zhenqin, et al. "MoleculeNet: a benchmark for molecular machine learning." *Chemical science* 9.2 (2018): 513-530.

1 Motivation & Problem Definition

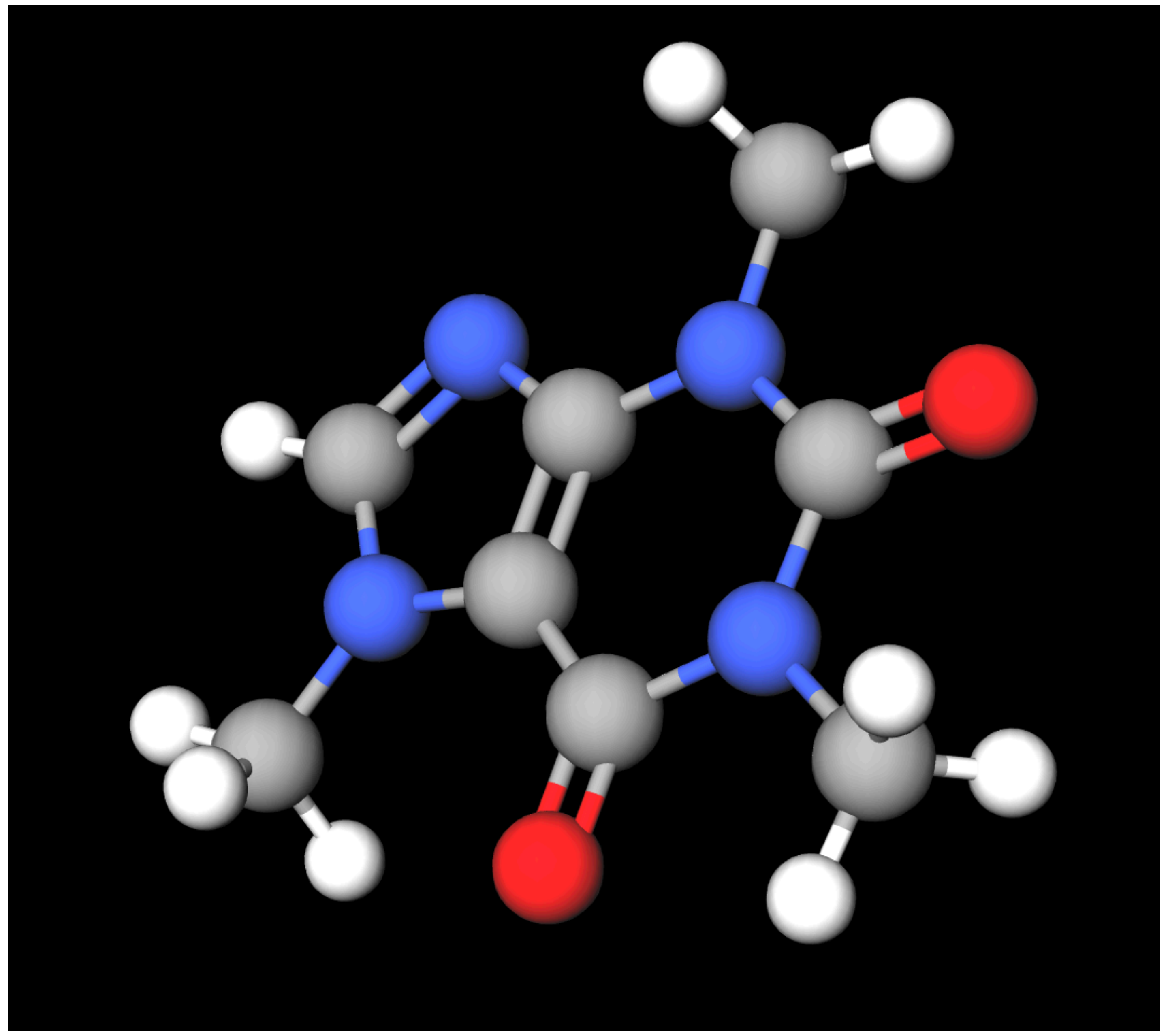
Indeed, molecules can also have 3D geometry.

1 Motivation & Problem Definition

Indeed, molecules can also have 3D geometry.



2D Molecular Graph

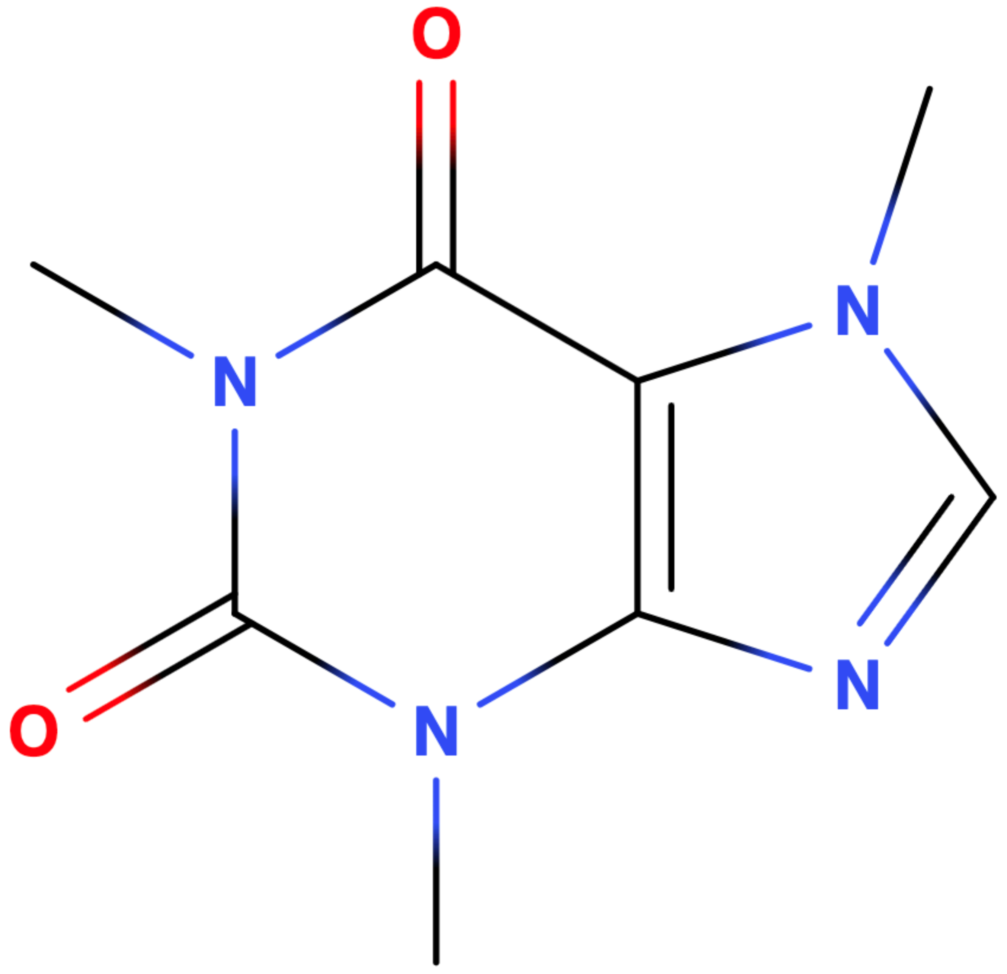


3D Molecular Graph

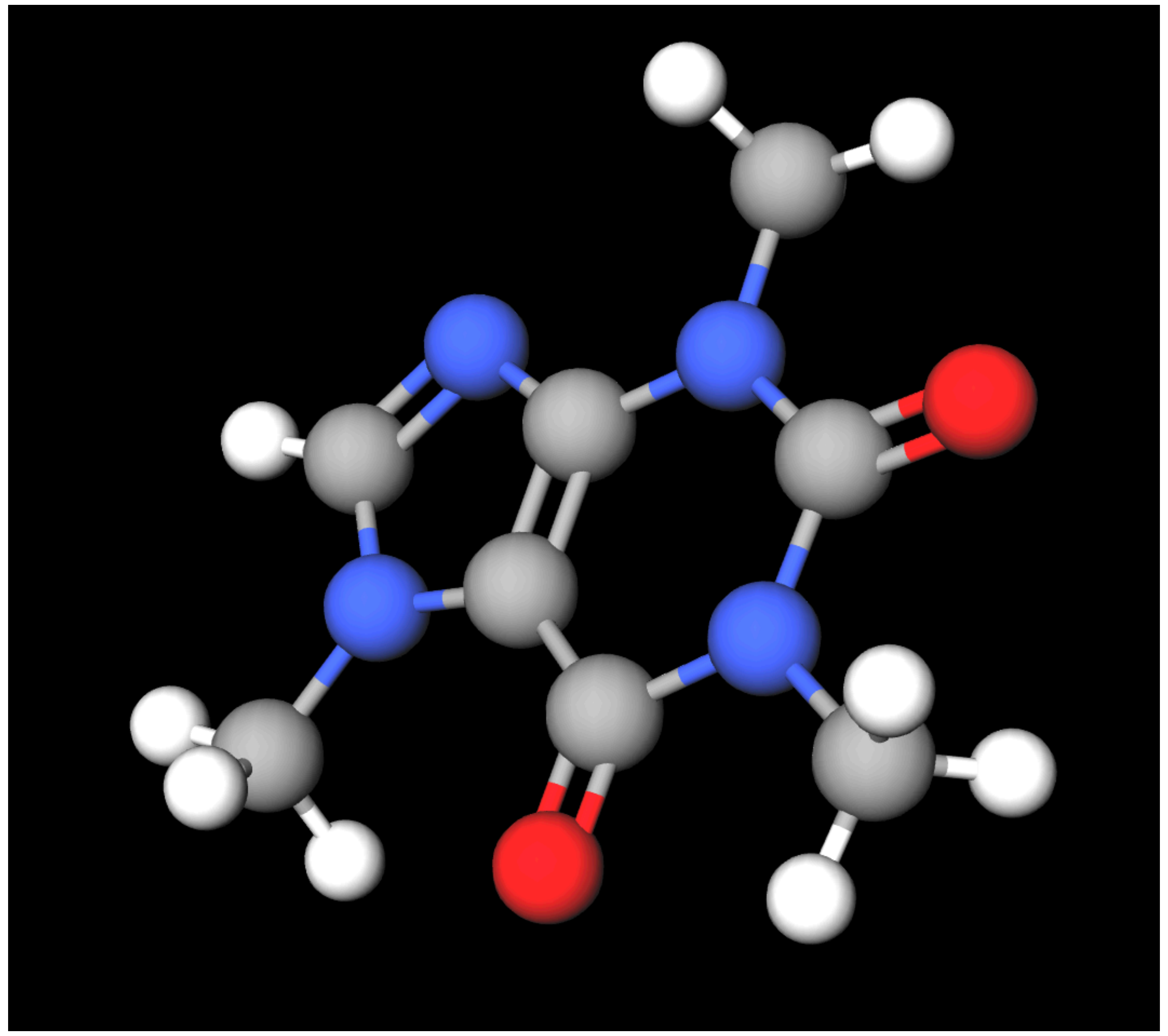
1 Motivation & Problem Definition

Indeed, molecules can also have 3D geometry.

- 3D geometry is more accurate for molecular property prediction.



2D Molecular Graph

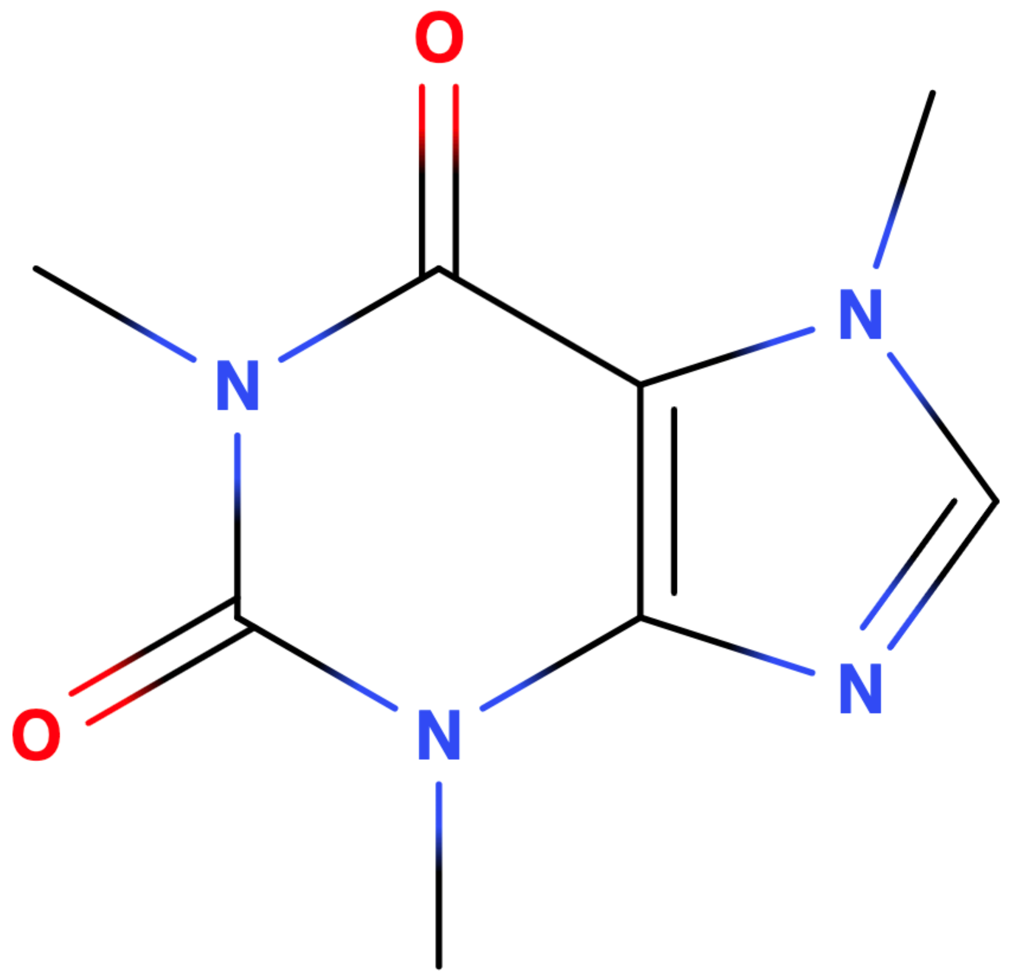


3D Molecular Graph

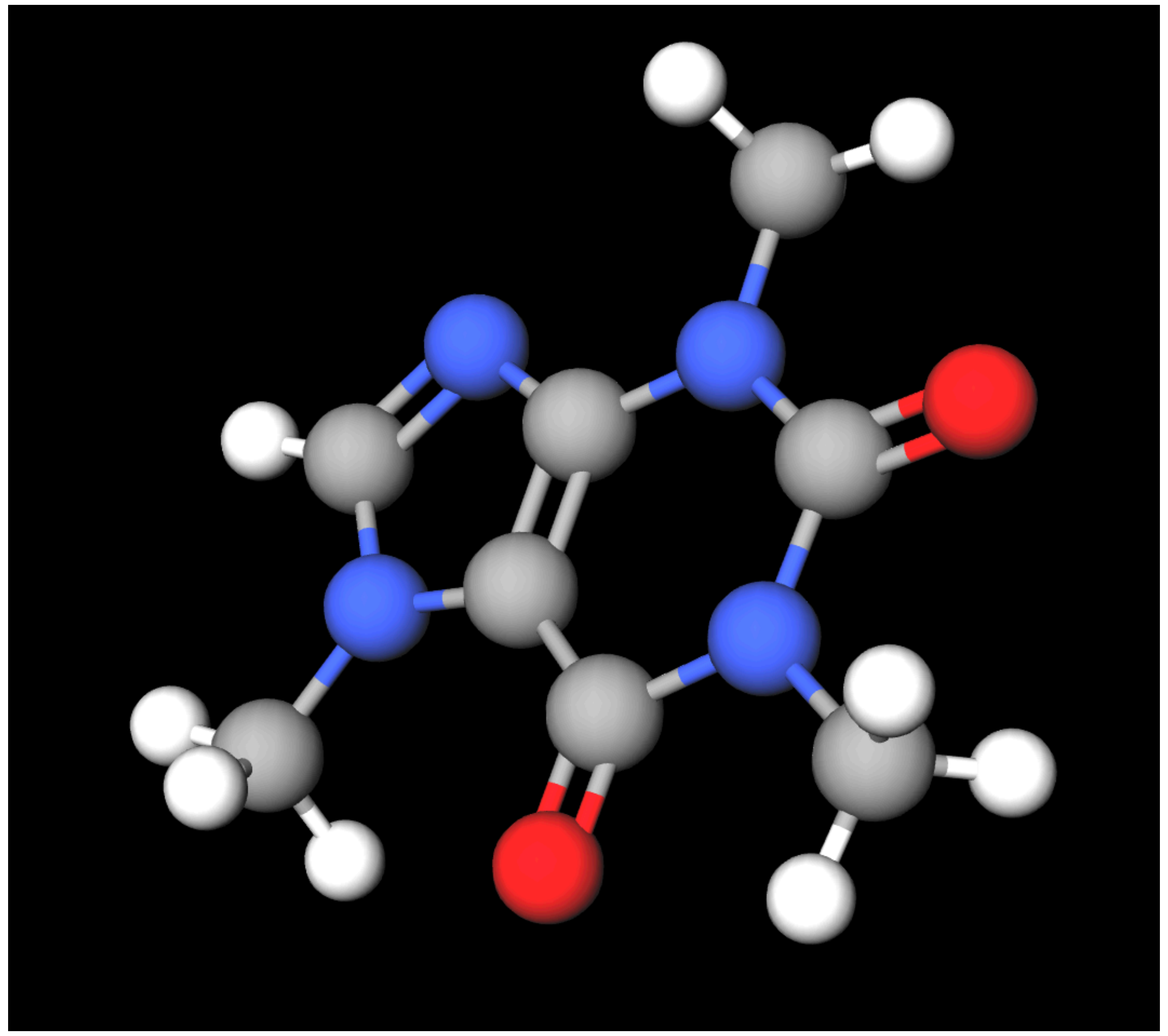
1 Motivation & Problem Definition

Indeed, molecules can also have 3D geometry.

- 3D geometry is more accurate for molecular property prediction.
- 3D geometry is more expensive to obtain (e.g. physical simulation).



2D Molecular Graph



3D Molecular Graph

1 Motivation & Problem Definition

Community has put more efforts in gathering large-scale 3D geometry datasets.
GEOM, Atom3D, Molecule3D, etc.

1 Motivation & Problem Definition

Community has put more efforts in gathering large-scale 3D geometry datasets.

GEOM, Atom3D, Molecule3D, etc.

Q: Can we utilize this for our ultimate goal?

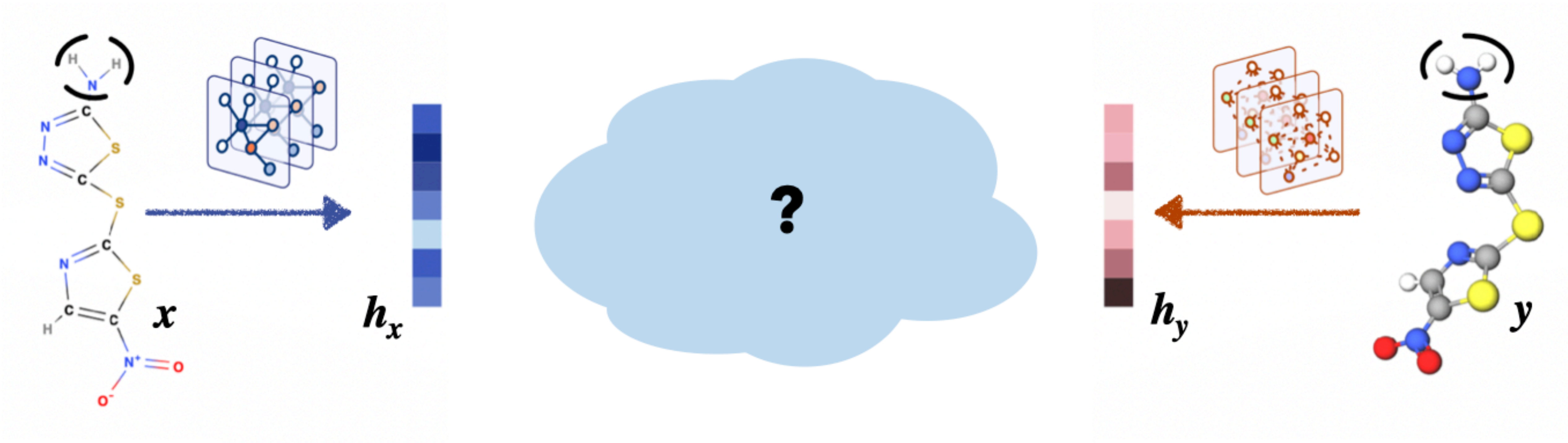
1 Motivation & Problem Definition

Community has put more efforts in gathering large-scale 3D geometry datasets.

GEOM, Atom3D, Molecule3D, etc.

Q: Can we utilize this for our ultimate goal?

A: Yes!



1 Motivation & Problem Definition

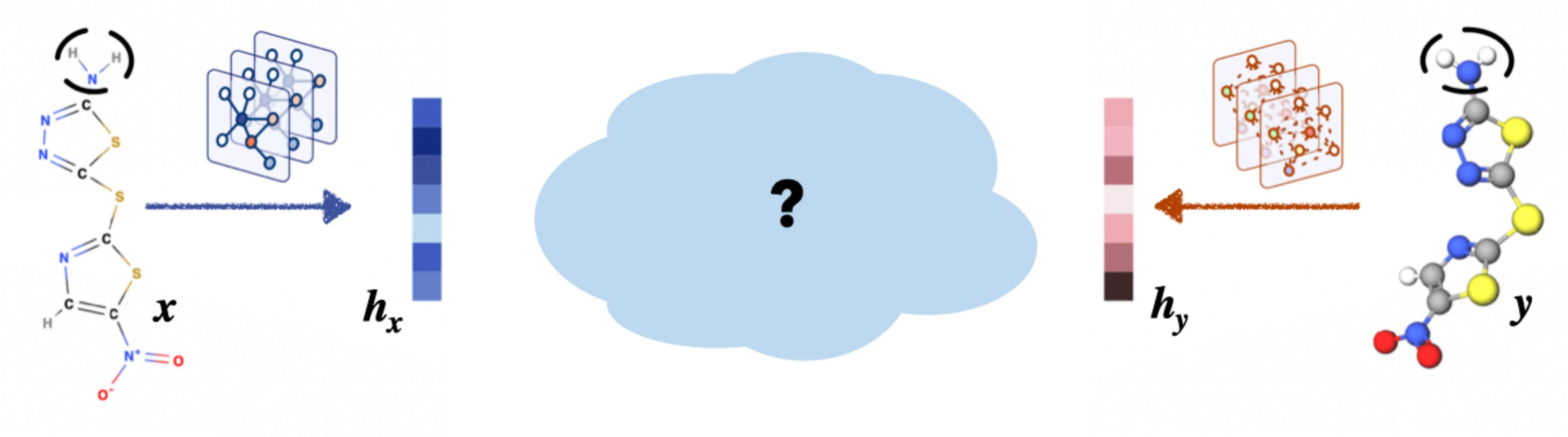
Community has put more efforts in gathering large-scale 3D geometry datasets.

GEOM, Atom3D, Molecule3D, etc.

Q: Can we utilize this for our ultimate goal?

A: Yes!

- **Graph Multi-View Pre-training (GraphMVP)** on 2D and 3D views.
- Pre-training: large-scale dataset with 2D and 3D graph.
- Fine-tuning: downstream tasks with 2D graph only.



1 Motivation & Problem Definition

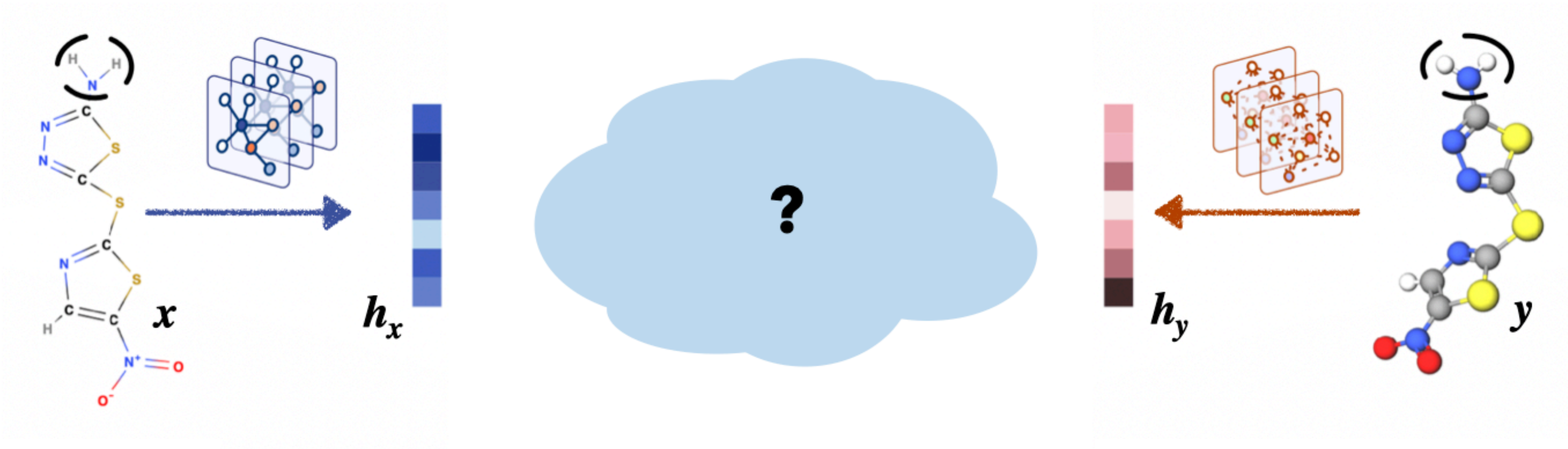
Community has put more efforts in gathering large-scale 3D geometry datasets.

GEOM, Atom3D, Molecule3D, etc.

Q: Can we utilize this for our ultimate goal?

A: Yes!

- **Graph Multi-View Pre-training (GraphMVP)** on 2D and 3D views.
- Pre-training: large-scale dataset with 2D and 3D graph.
- Fine-tuning: downstream tasks with 2D graph only.



General:

- **Two augmentation views in SimCLR.**
- **Local and global views in Deep InfoMax.**
- **Masked and visual patches in BEiT.**
- **…**

2 Related Work

Contrastive and Generative SSL have been widely discussed in [1, 2, 3, 4].

[1] Liu, Xiao, et al. "Self-supervised learning: Generative or contrastive." *IEEE Transactions on Knowledge and Data Engineering* (2021).
[2] Liu, Yixin, et al. "Graph self-supervised learning: A survey." *arXiv preprint arXiv:2103.00111* (2021).
[3] Wu, Lirong, et al. "Self-supervised on graphs: Contrastive, generative, or predictive." *arXiv e-prints* (2021): arXiv-2105.
[4] Xie, Yaochen, et al. "Self-supervised learning of graph neural networks: A unified review." *arXiv preprint arXiv:2102.10757* (2021).

2 Related Work

Contrastive and Generative SSL have been widely discussed in [1, 2, 3, 4].

Contrastive SSL:

- Inter-data
- Examples: InfoNCE, Jense-Shannon Estimation

[1] Liu, Xiao, et al. "Self-supervised learning: Generative or contrastive." *IEEE Transactions on Knowledge and Data Engineering* (2021).
[2] Liu, Yixin, et al. "Graph self-supervised learning: A survey." *arXiv preprint arXiv:2103.00111* (2021).
[3] Wu, Lirong, et al. "Self-supervised on graphs: Contrastive, generative, or predictive." *arXiv e-prints* (2021): arXiv-2105.
[4] Xie, Yaochen, et al. "Self-supervised learning of graph neural networks: A unified review." *arXiv preprint arXiv:2102.10757* (2021).

2 Related Work

Contrastive and Generative SSL have been widely discussed in [1, 2, 3, 4].

Contrastive SSL:

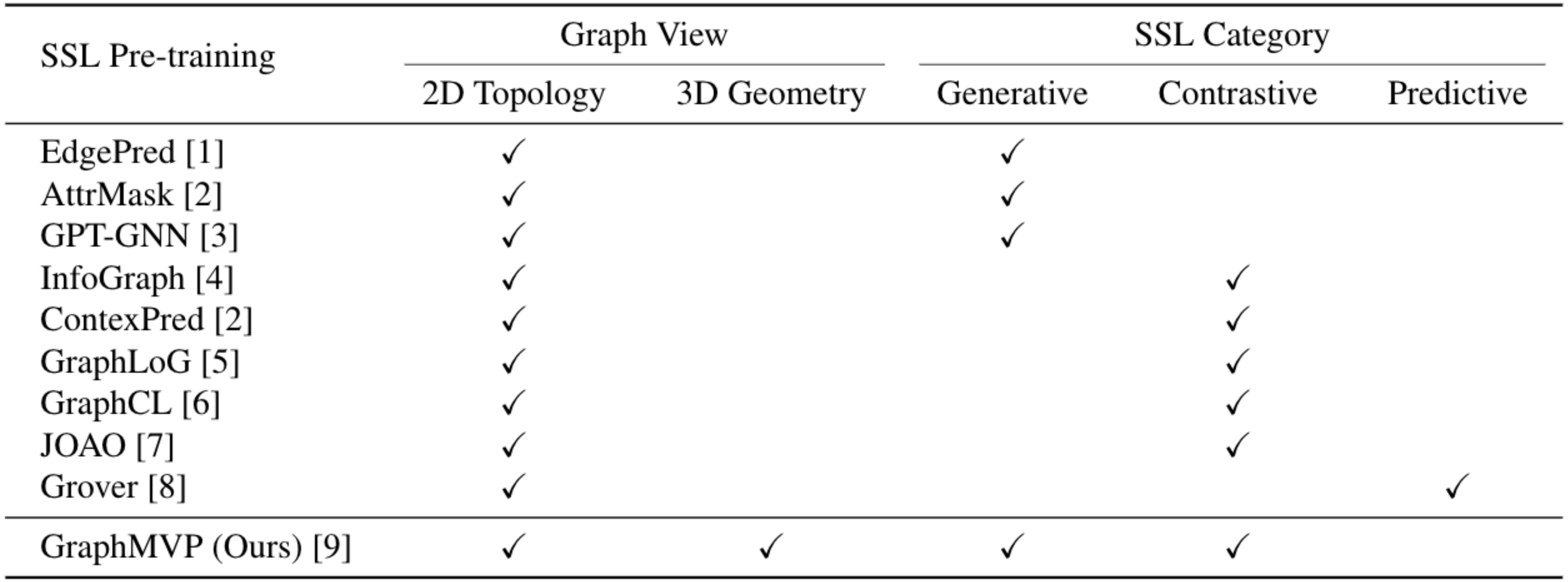
- Inter-data
- Examples: InfoNCE, Jense-Shannon Estimation

Generative SSL:

- Intra-data
- Examples: Masked Auto-Encoding, BYOL, SimSiam

[1] Liu, Xiao, et al. "Self-supervised learning: Generative or contrastive." *IEEE Transactions on Knowledge and Data Engineering* (2021).
[2] Liu, Yixin, et al. "Graph self-supervised learning: A survey." *arXiv preprint arXiv:2103.00111* (2021).
[3] Wu, Lirong, et al. "Self-supervised on graphs: Contrastive, generative, or predictive." *arXiv e-prints* (2021): arXiv-2105.
[4] Xie, Yaochen, et al. "Self-supervised learning of graph neural networks: A unified review." *arXiv preprint arXiv:2102.10757* (2021).

2 Related Work



SSL Pre-training	Graph View		SSL Category		
	2D Topology	3D Geometry	Generative	Contrastive	Predictive
EdgePred [1]	✓		✓		
AttrMask [2]	✓		✓		
GPT-GNN [3]	✓		✓		
InfoGraph [4]	✓			✓	
ContexPred [2]	✓			✓	
GraphLoG [5]	✓			✓	
GraphCL [6]	✓			✓	
JOAO [7]	✓			✓	
Grover [8]	✓				✓
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 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, $h_x =$ 2D-GNN(A, E).

Molecule as 3D geometric graph:

- y for a 3D molecular graph.
- h_y for 3D representation, $h_y =$ 3D-GNN(A, R).

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).

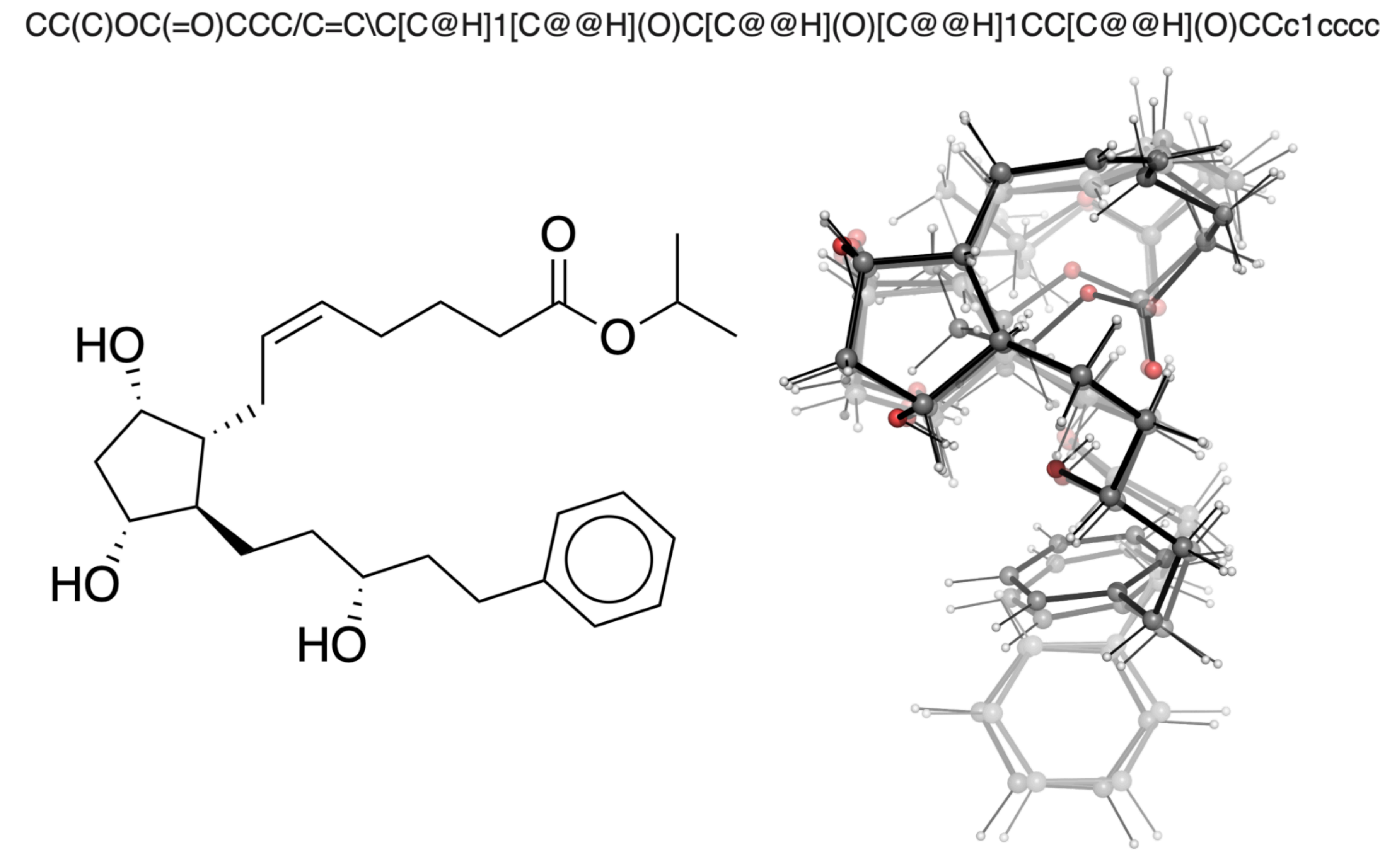

CC(C)OC(=O)CCC/C=C\C[C@H]1[C@@H](O)C[C@@H](O)[C@@H]1CC[C@@H](O)CCc1ccccc1
O
O
HO
HO
HO

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 Preliminaries

Energy-Based Model (EBM): $p(x) = \frac{\exp(-E(x))}{A}$, where $E(x)$ is the energy function, and $A = \int_x \exp(-E(x))dx$ is normalization constant / partition function.

3 Preliminaries

Energy-Based Model (EBM): $p(x) = \dfrac{\exp(-E(x))}{A}$, where $E(x)$ is the energy function, and $A = \displaystyle\int_x \exp(-E(x))dx$ is normalization constant / partition function.

- Bottleneck: intractable A

3 Preliminaries

Energy-Based Model (EBM): $p(x) = \frac{\exp(-E(x))}{A}$, where $E(x)$ is the energy function, and $A = \int_x \exp(-E(x))dx$ is normalization constant / partition function.

- Bottleneck: intractable A
- 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.

4 Method: GraphMVP

4.1 MI and SSL

4.2 Contrastive SSL

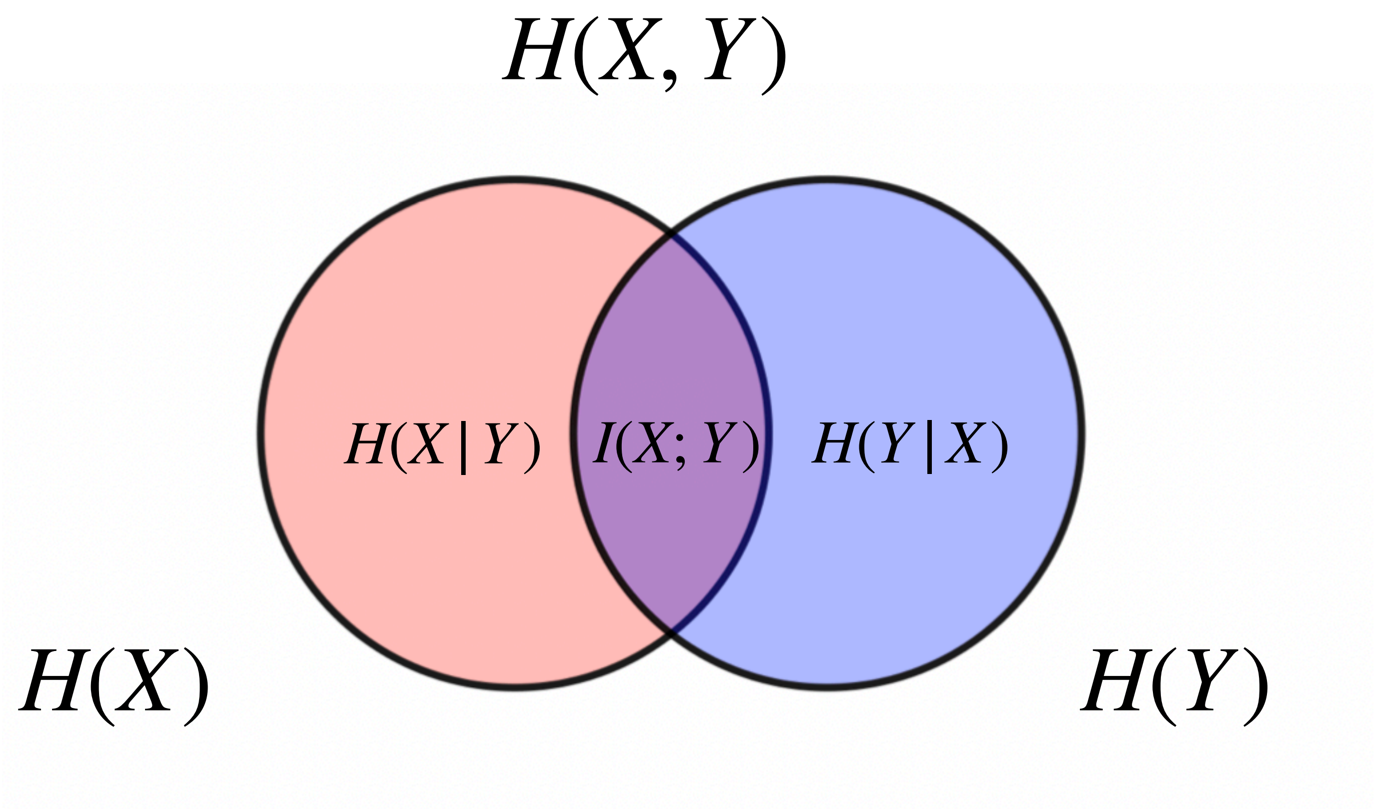
4.3 Generative SSL

4.4 Multi-task Objective

4.1 MI and SSL

Mutual information (MI):

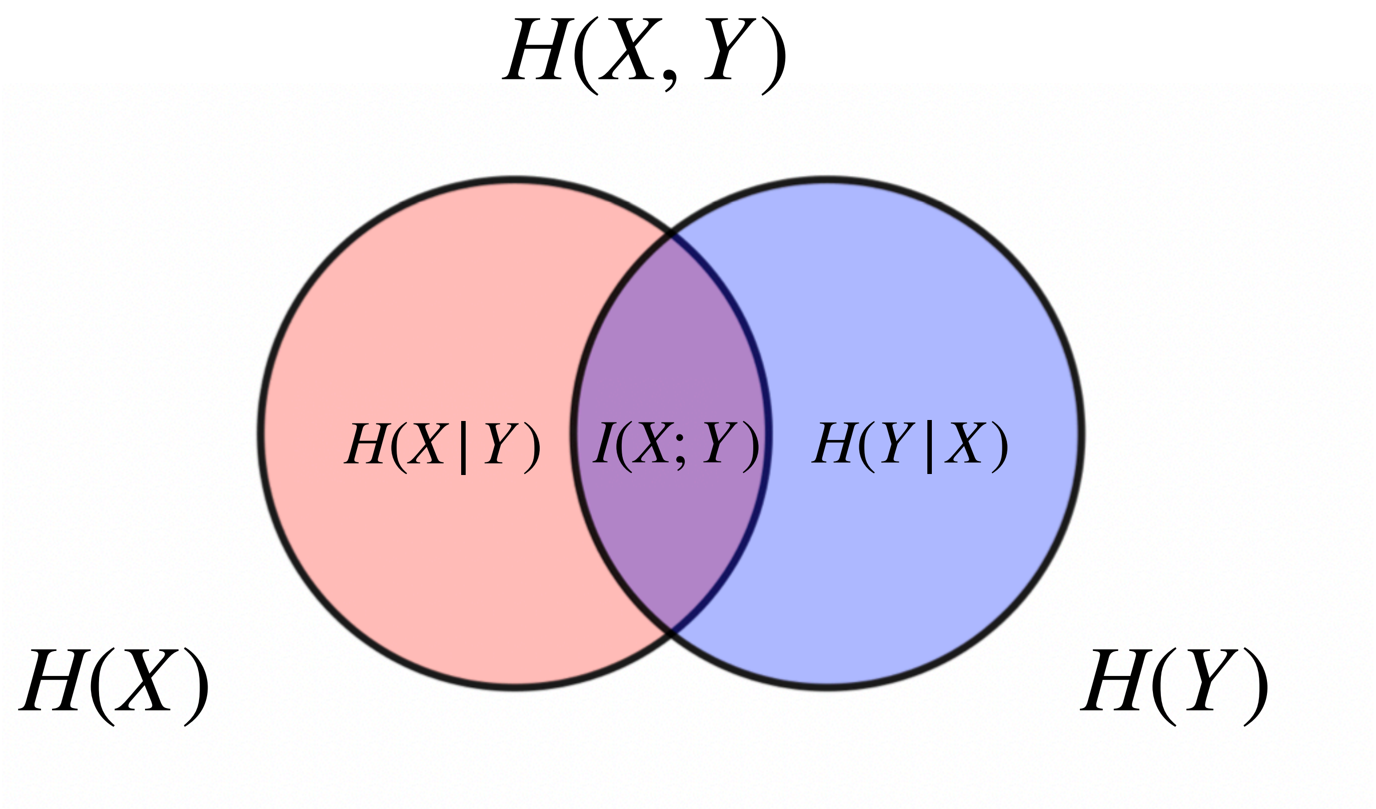
- measures the non-linear dependence between variables.
- the larger MI, the stronger dependence between variables.



4.1 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 expressive 2D representation by sharing more information with its 3D counterparts.

4.1 MI and SSL

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

How to maximize this?

4.1 MI and SSL

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

How to maximize this?

GraphMVP proposes 2 SSL frameworks — 1 contrastive and 1 generative.

4.2 Contrastive SSL

Lower bound on MI:

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

4.2 Contrastive SSL

Lower bound on MI:

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

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

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

4.2 Contrastive SSL

Lower bound on MI:

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

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

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

What does this mean?

4.2 Contrastive SSL

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

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

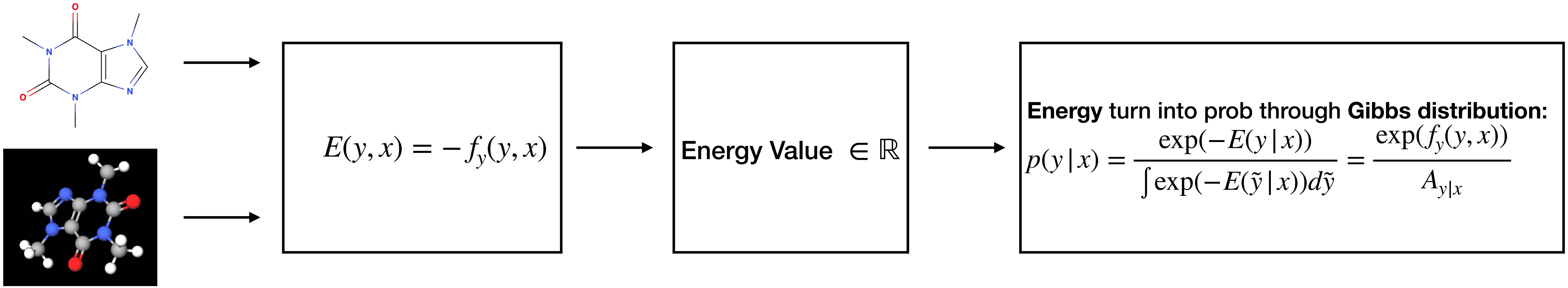
What does this mean?

4.2 Contrastive SSL

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

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

What does this mean?



$$E(y,x) = -f_y(y,x)$$

Energy Value $\in \mathbb{R}$

Energy turn into prob through **Gibbs distribution:**

$$p(y\,|\,x) = \frac{\exp(-E(y\,|\,x))}{\int \exp(-E(\tilde{y}\,|\,x))d\tilde{y}} = \frac{\exp(f_y(y,x))}{A_{y|x}}$$

4.2 Contrastive SSL

Lower bound on MI:

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

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

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

4.2 Contrastive SSL

Lower bound on MI:

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

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

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

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

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

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

4.2 Contrastive SSL

EBM-NCE & Jensen-Shannon Estimation (JSE)

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

4.2 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.

4.2 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.

[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).

4.2 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).

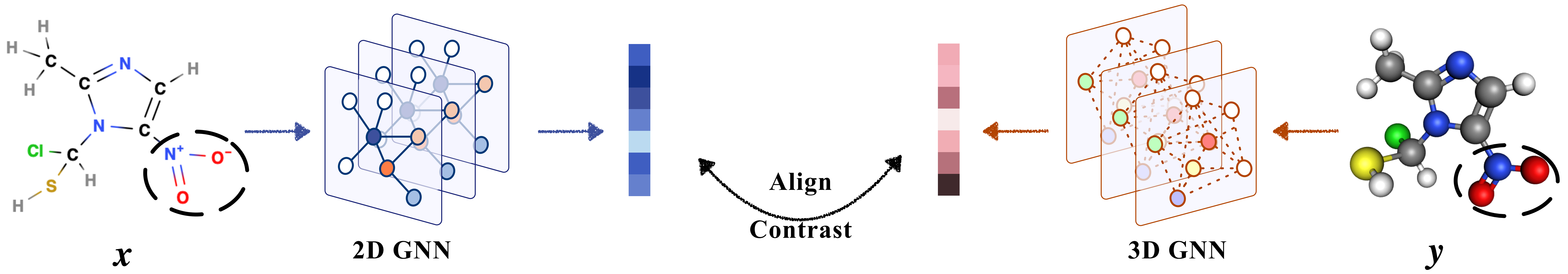
4.2 Contrastive SSL

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.,

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



4.3 Generative SSL

Lower bound on MI:

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

4.3 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 $z_x \sim \mathcal{N}(z_x; \mu_x, \Sigma_x)$:

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

4.3 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 $z_x \sim \mathcal{N}(z_x; \mu_x, \Sigma_x)$:

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

Reconstruction

4.3 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).

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

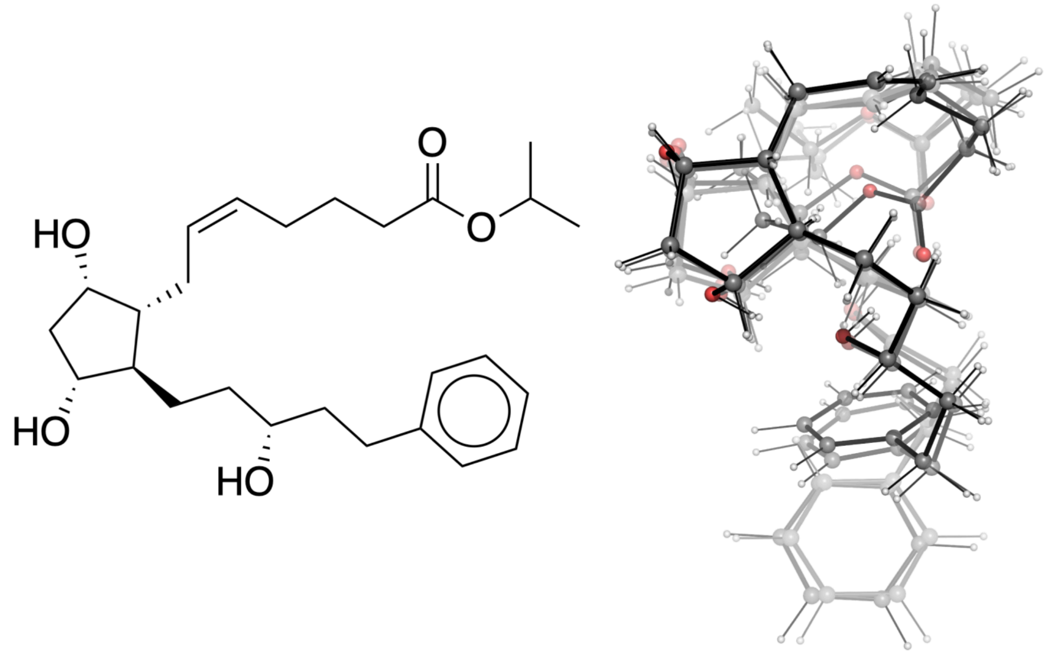


Figure 1. Molecular representations of the latanoprost molecule. *top* SMILES string. *left* Stereochemical formula with edge features, including wedges for in- and out-of-plane bonds, and a double line for *cis* isomerism. *right* Overlay of conformers. Higher transparency corresponds to lower statistical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

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 $z_x \sim \mathcal{N}(z_x; \mu_x, \Sigma_x)$:

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

Benefits:

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

4.3 Generative SSL

Lower bound on MI:

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

Variational Molecule Reconstruction

We introduce a variational distribution $z_x \sim \mathcal{N}(z_x; \mu_x, \Sigma_x)$:

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

Reconstruction

Limitation:

- Decoder for structured data. If the target data space, like 3D and 2D molecule, is discrete/structured, then the modeling and evaluation on this data space is hard.

4.3 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 $z_x \sim \mathcal{N}(z_x; \mu_x, \Sigma_x)$:

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

Reconstruction

Solution:

Variational Representation Reconstruction (VRR)

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

4.3 Generative SSL

Variational Molecule Reconstruction

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

Variational Representation Reconstruction

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

4.3 Generative SSL

Variational Molecule Reconstruction

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

Variational Representation Reconstruction

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

Recall: 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.

4.3 Generative SSL

Variational Molecule Reconstruction

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

Variational Representation Reconstruction

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

1. 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.

4.3 Generative SSL

Variational Molecule Reconstruction

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

Variational Representation Reconstruction

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

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

4.3 Generative SSL

Variational Molecule Reconstruction

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

Variational Representation Reconstruction

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

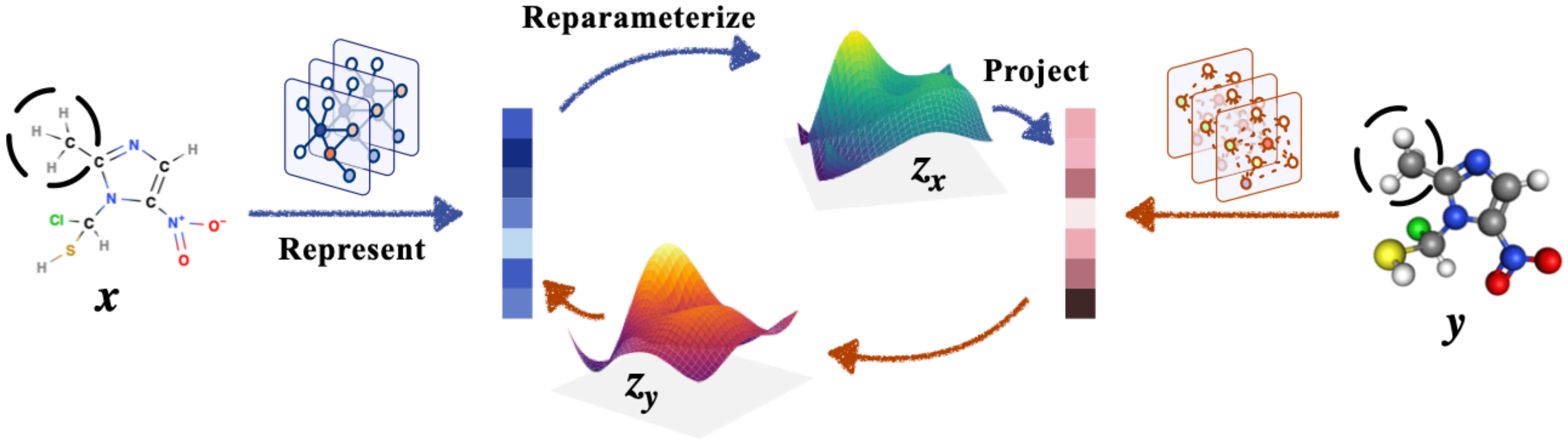
1. If is y is discrete and structured, then we propose this surrogate loss: $\|h_y(y) - h_y(g_x(z_x))\|^2$, where h_y is the encoder on y
2. By approximation: $\|h_y(y) - q_x(z_x))\|^2$
3. Add stop-gradient: $\|\boxed{\text{SG}}(h_y(y)) - q_x(z_x))\|^2$

4.3 Generative SSL

Final solution (VRR):

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

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

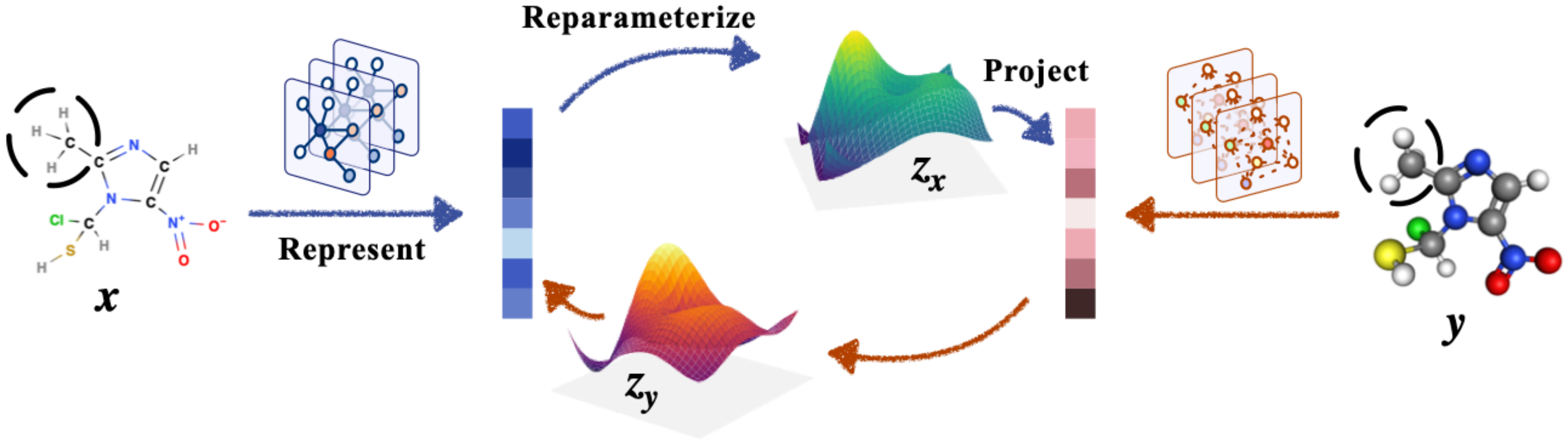


4.3 Generative SSL

Final solution (VRR):

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

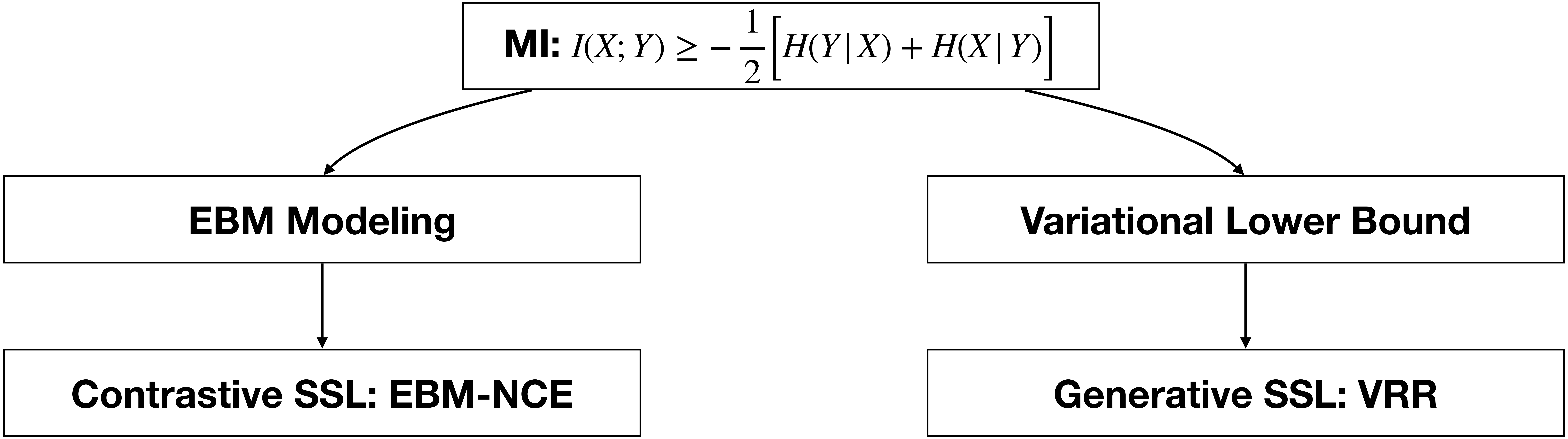
$$+\frac{\beta}{2}\cdot\Big[KL(q(z_x|x)||p(z_x)) + KL(q(z_y|y)||p(z_y))\Big].$$



Notice 1: this surrogate loss can be exact if h_x/h_y is continuous invertible.

Notice 2: this is another form of non-contrastive SSL (BYOL/SimSiam).

4.4 Multi-task Objective



MI: $I(X;Y) \geq -\frac{1}{2}\Big[H(Y|X) + H(X|Y)\Big]$

EBM Modeling

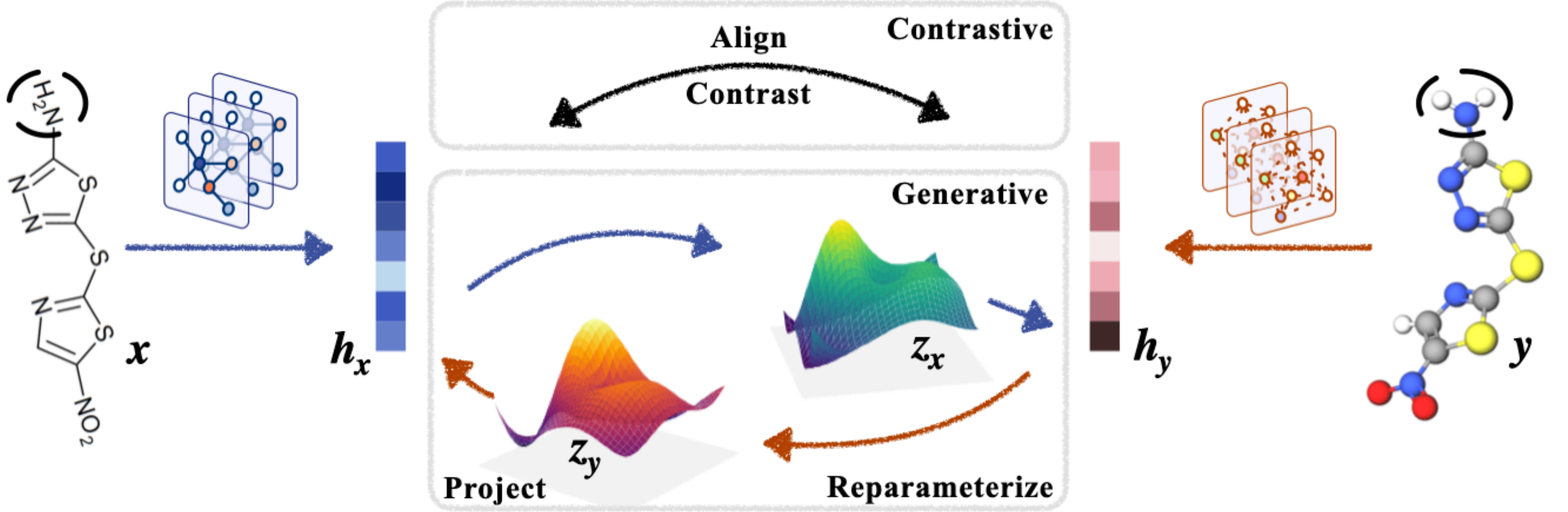
Variational Lower Bound

Contrastive SSL: EBM-NCE

Generative SSL: VRR

4.4 Multi-task Objective

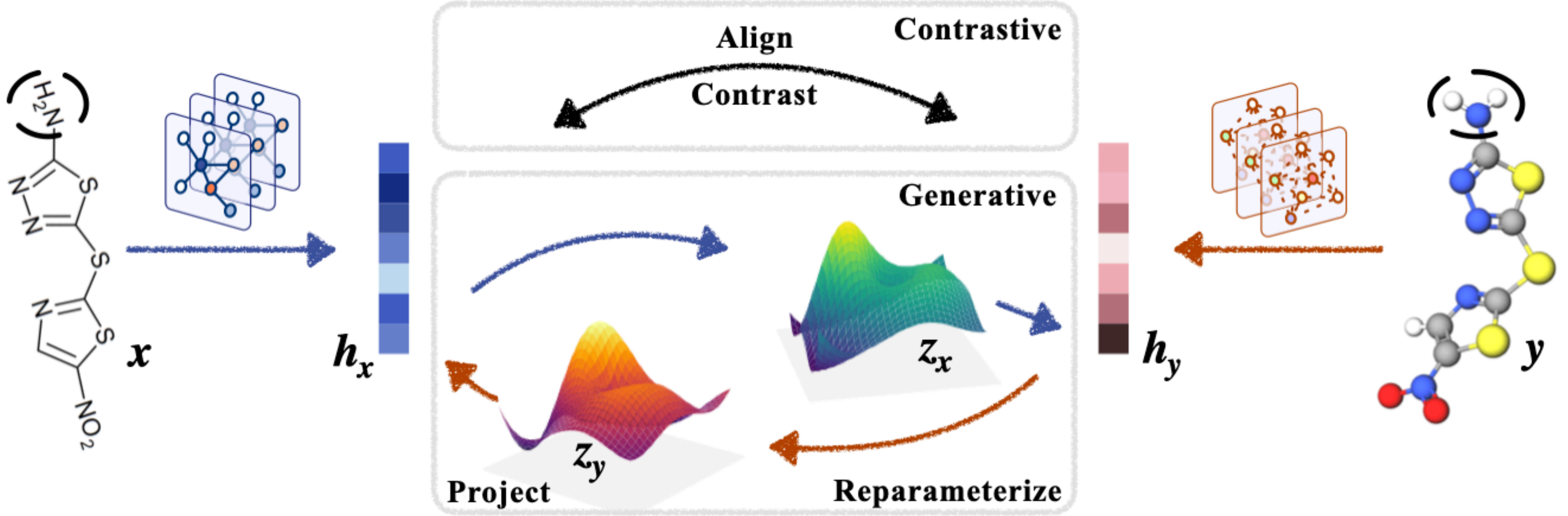
The objective is weighted sum of the contrastive and generative SSL:
$\mathscr{L}_{\text{GraphMVP}} = \alpha_1 \cdot \mathscr{L}_{\text{C}} + \alpha_2 \cdot \mathscr{L}_{\text{G}}$.



4.4 Multi-task Objective

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

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



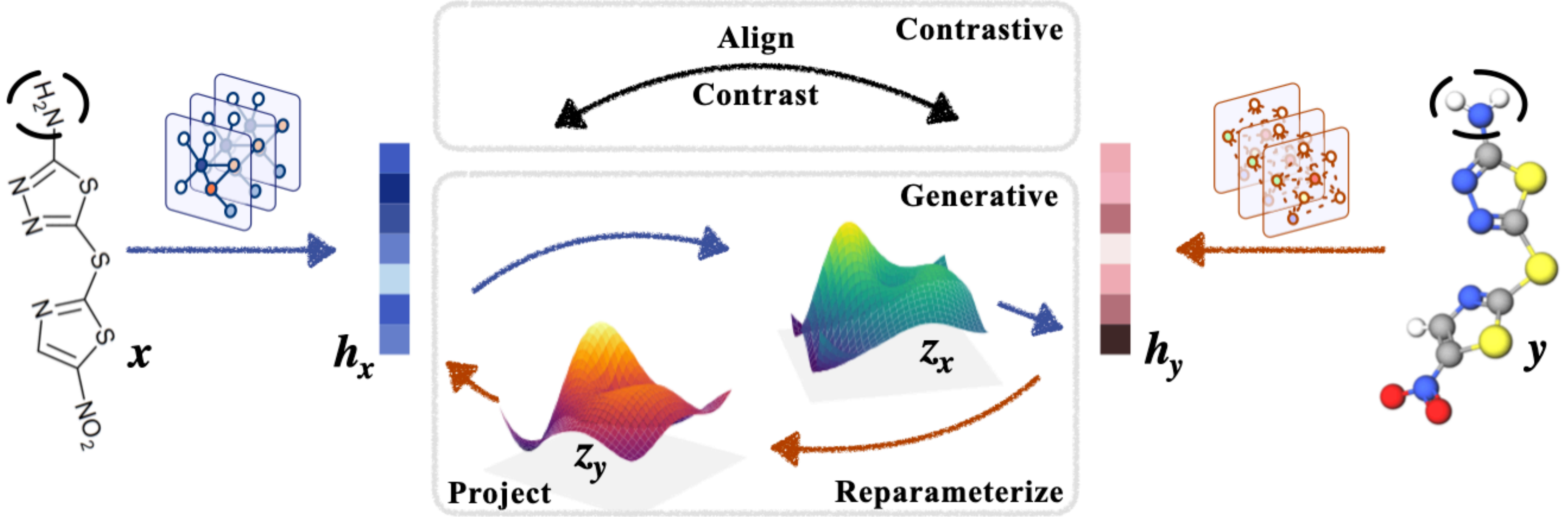
Contrastive and generative SSL are complementary.

- 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.

4.4 Multi-task Objective

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

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



Till now, GraphMVP considers only doing SSL between 3D and 2D views; yet the 2D SSL can also be merged with it:

- $\mathscr{L}_{\text{GraphMVP-G}} = \mathscr{L}_{\text{GraphMVP}} + \alpha_3 \cdot \mathscr{L}_{\text{Generative 2D-SSL}}$
- $\mathscr{L}_{\text{GraphMVP-C}} = \mathscr{L}_{\text{GraphMVP}} + \alpha_3 \cdot \mathscr{L}_{\text{Contrastive 2D-SSL}}$

5 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.

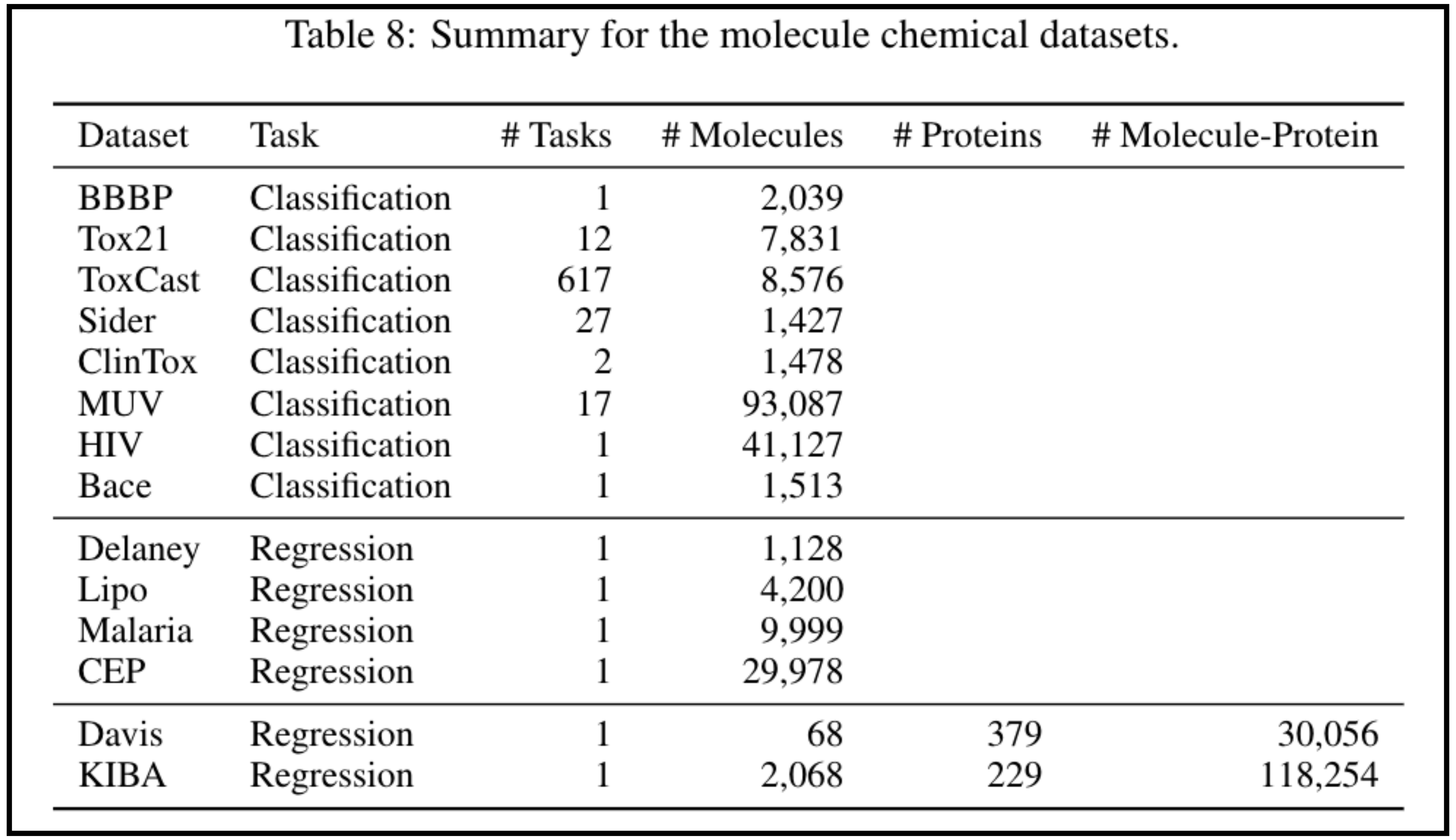


Table 8: Summary for the molecule chemical datasets.

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.

5 Experiments

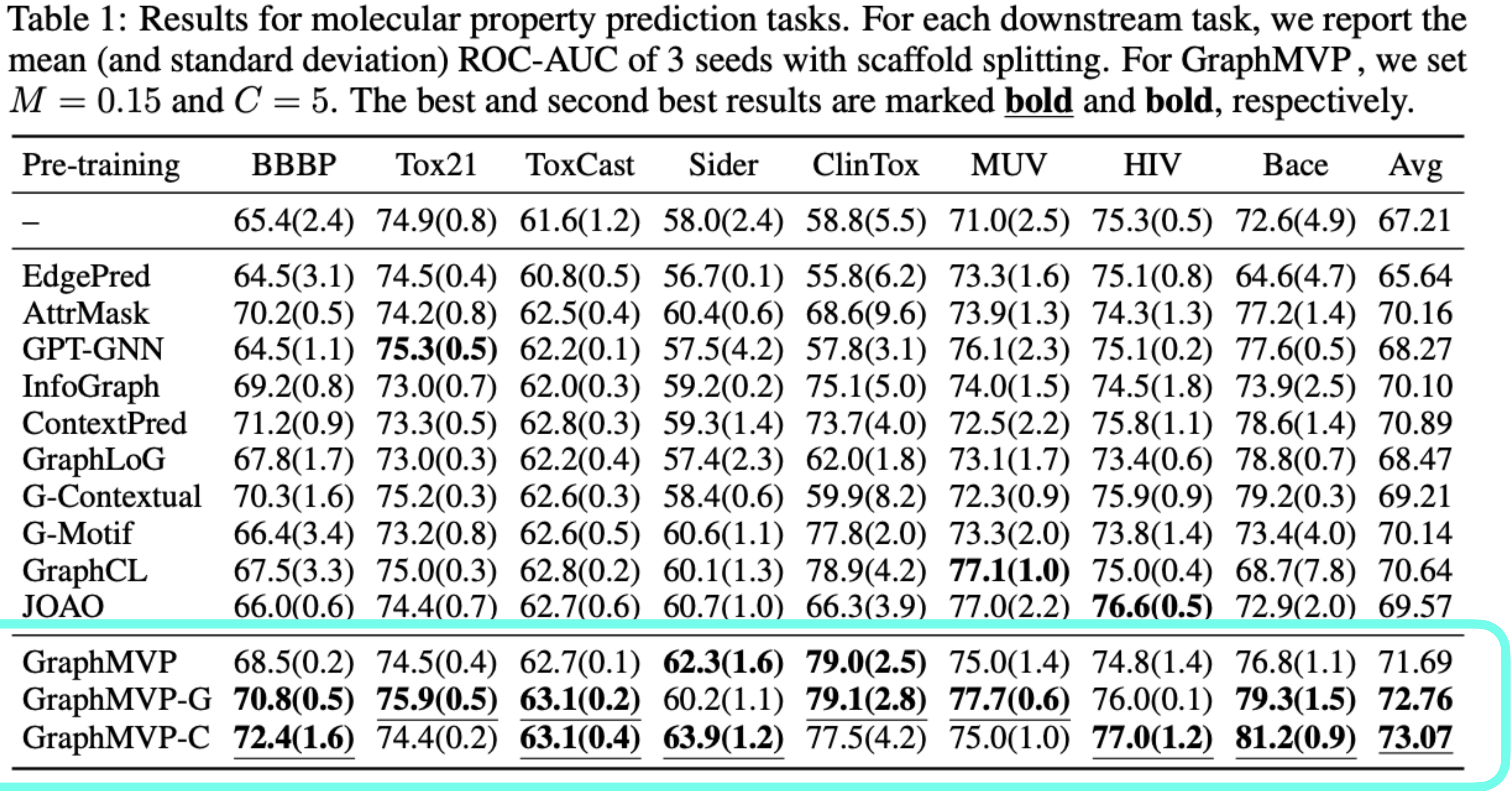


Table 1: Results for molecular property prediction tasks. For each downstream task, we report the mean (and standard deviation) ROC-AUC of 3 seeds with scaffold splitting. For GraphMVP , we set $M = 0.15$ and $C = 5$. The best and second best results are marked **bold** and **bold**, respectively.

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**

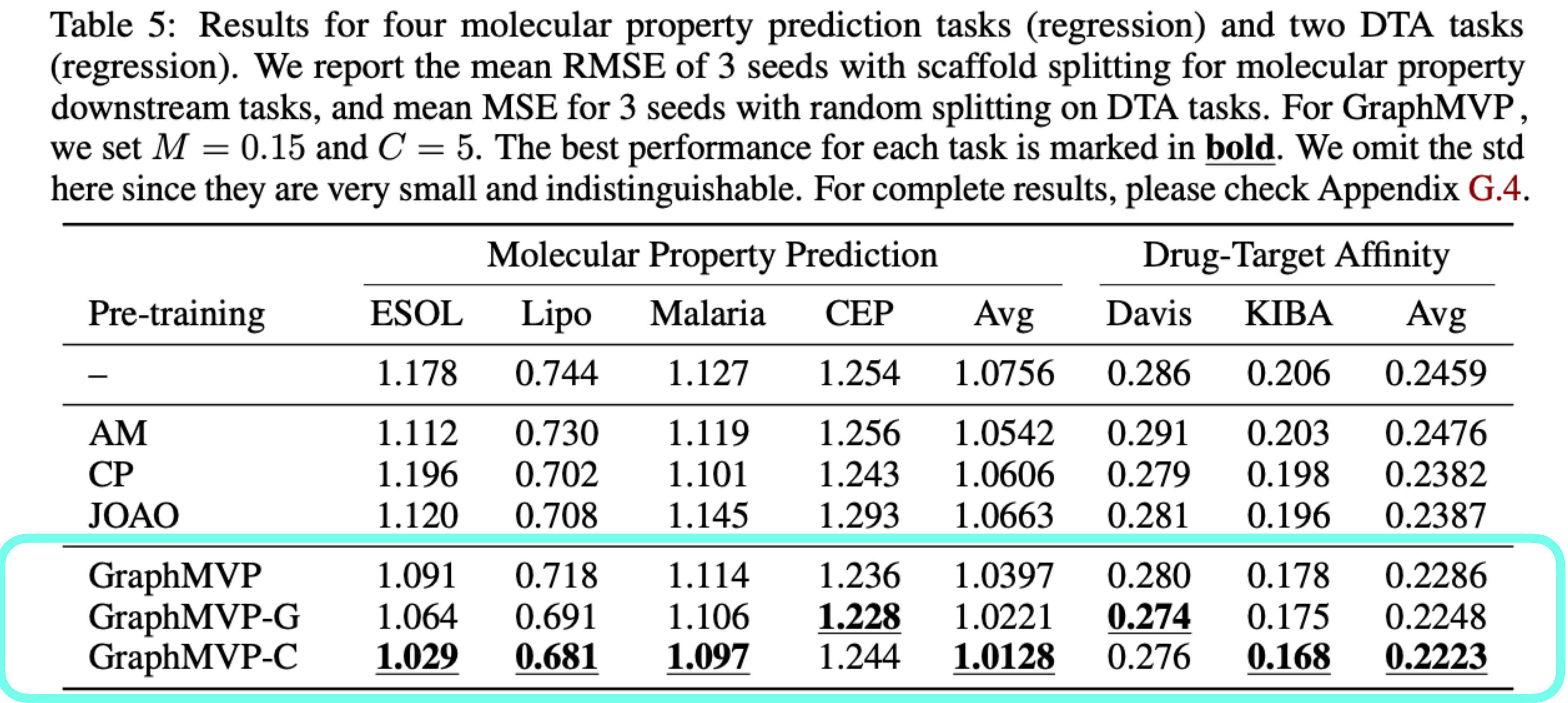
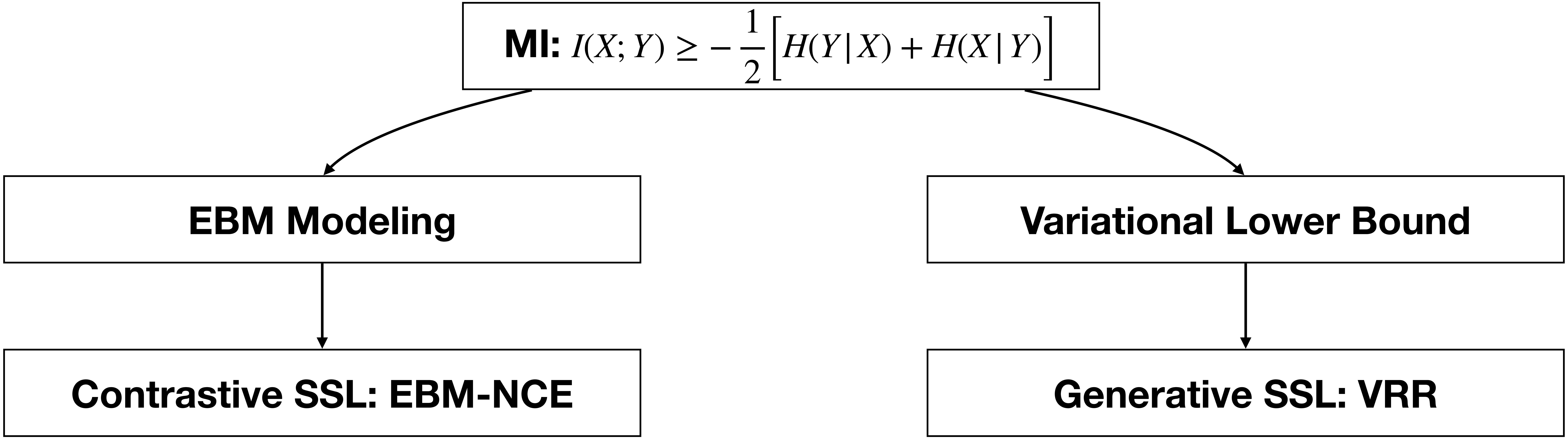


Table 5: Results for four molecular property prediction tasks (regression) and two DTA tasks (regression). We report the mean RMSE of 3 seeds with scaffold splitting for molecular property downstream tasks, and mean MSE for 3 seeds with random splitting on DTA tasks. For GraphMVP , we set $M = 0.15$ 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**

6 Future Directions: SSL on Structured Data



MI: $I(X;Y) \geq -\frac{1}{2}\Big[H(Y|X) + H(X|Y)\Big]$

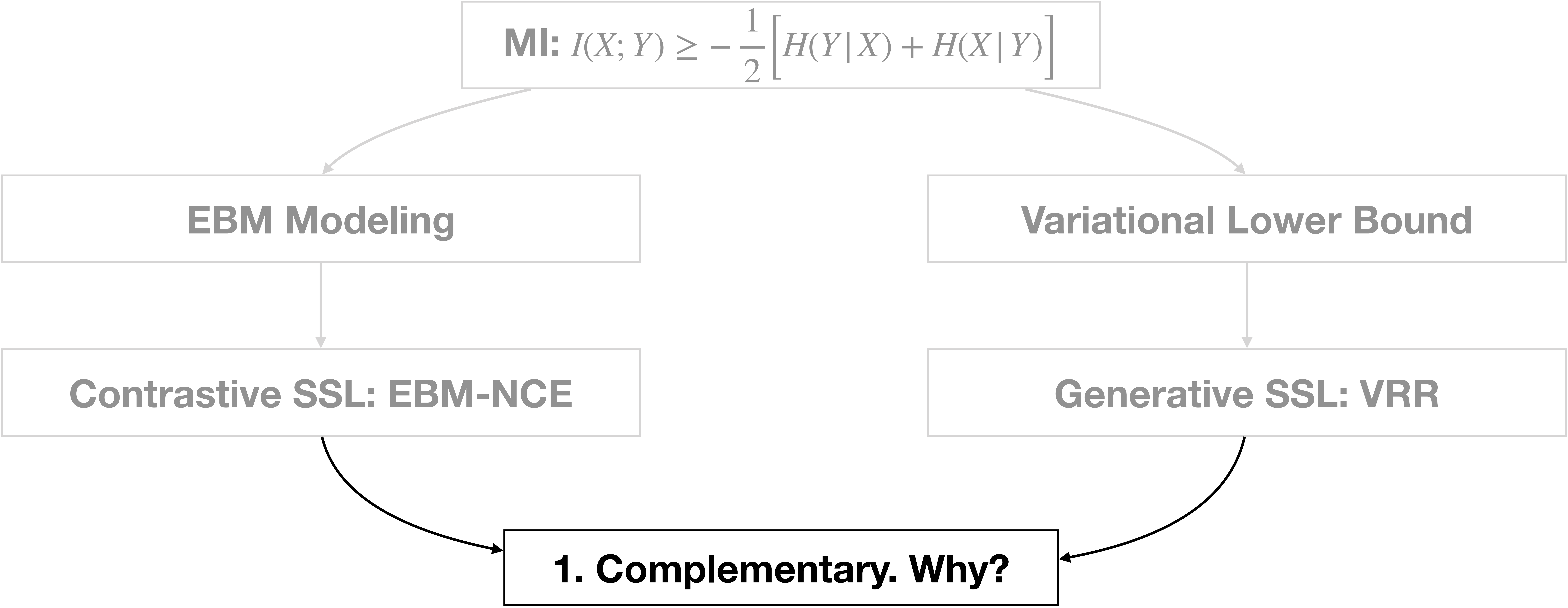
EBM Modeling

Variational Lower Bound

Contrastive SSL: EBM-NCE

Generative SSL: VRR

6 Future Directions: SSL on Structured Data



MI: $I(X;Y) \geq -\frac{1}{2}\left[H(Y|X) + H(X|Y)\right]$

EBM Modeling

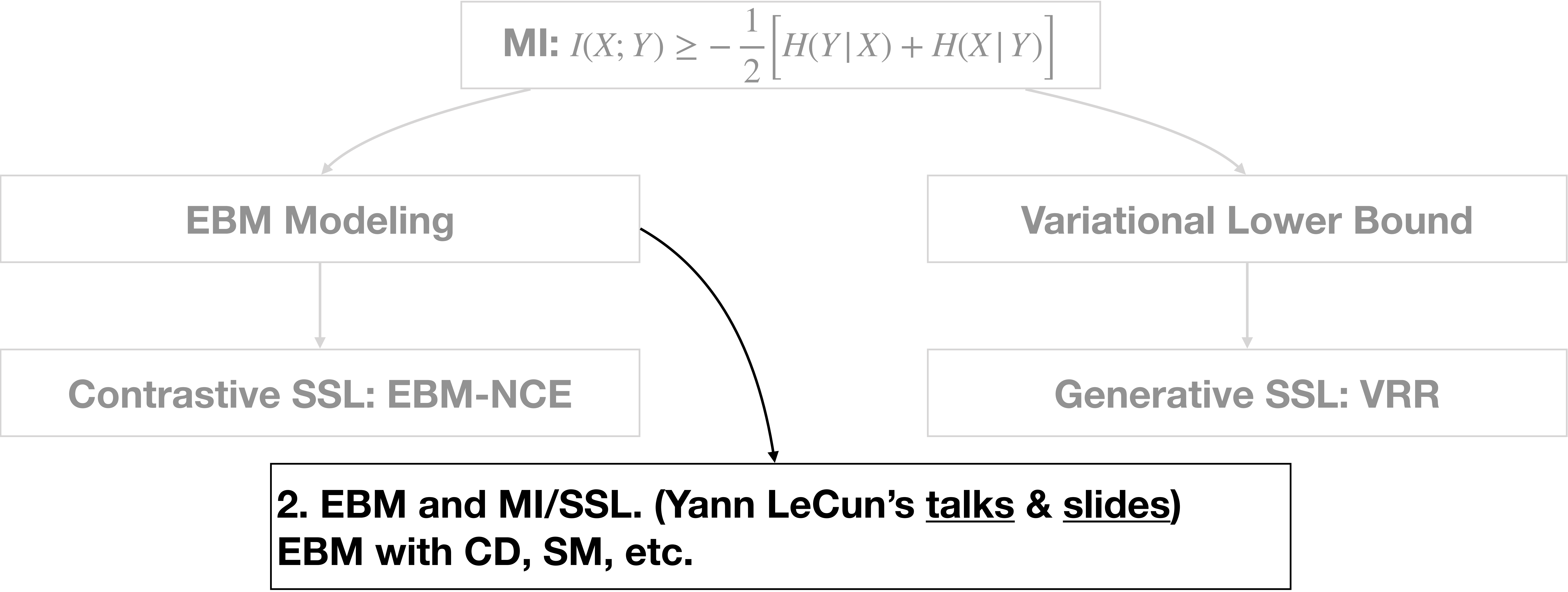
Variational Lower Bound

Contrastive SSL: EBM-NCE

Generative SSL: VRR

1. Complementary. Why?

6 Future Directions: SSL on Structured Data



MI: $I(X;Y) \geq -\frac{1}{2}\Big[H(Y|X) + H(X|Y)\Big]$

EBM Modeling

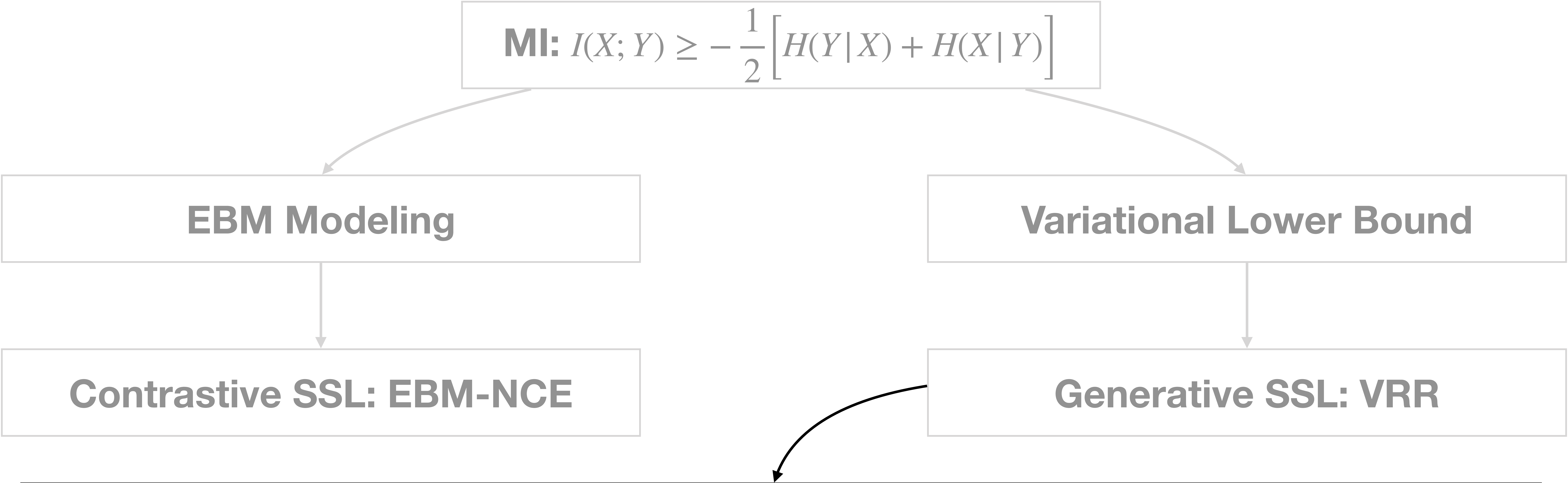
Variational Lower Bound

Contrastive SSL: EBM-NCE

Generative SSL: VRR

2. EBM and MI/SSL. (Yann LeCun's talks & slides)
EBM with CD, SM, etc.

6 Future Directions: SSL on Structured Data



MI: $I(X;Y) \geq -\frac{1}{2}\left[H(Y|X) + H(X|Y)\right]$

EBM Modeling

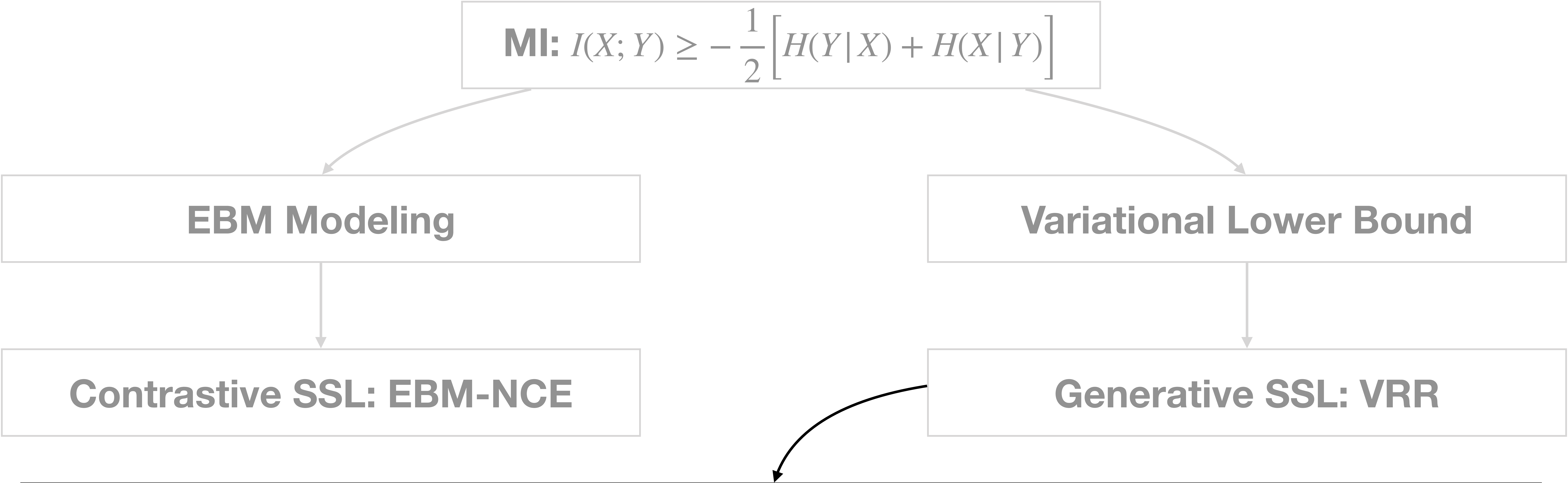
Variational Lower Bound

Contrastive SSL: EBM-NCE

Generative SSL: VRR

3. Another way to understand non-contrastive SSL.

6 Future Directions: SSL on Structured Data



MI: $I(X;Y) \geq -\frac{1}{2}\Big[H(Y|X) + H(X|Y)\Big]$

EBM Modeling

Variational Lower Bound

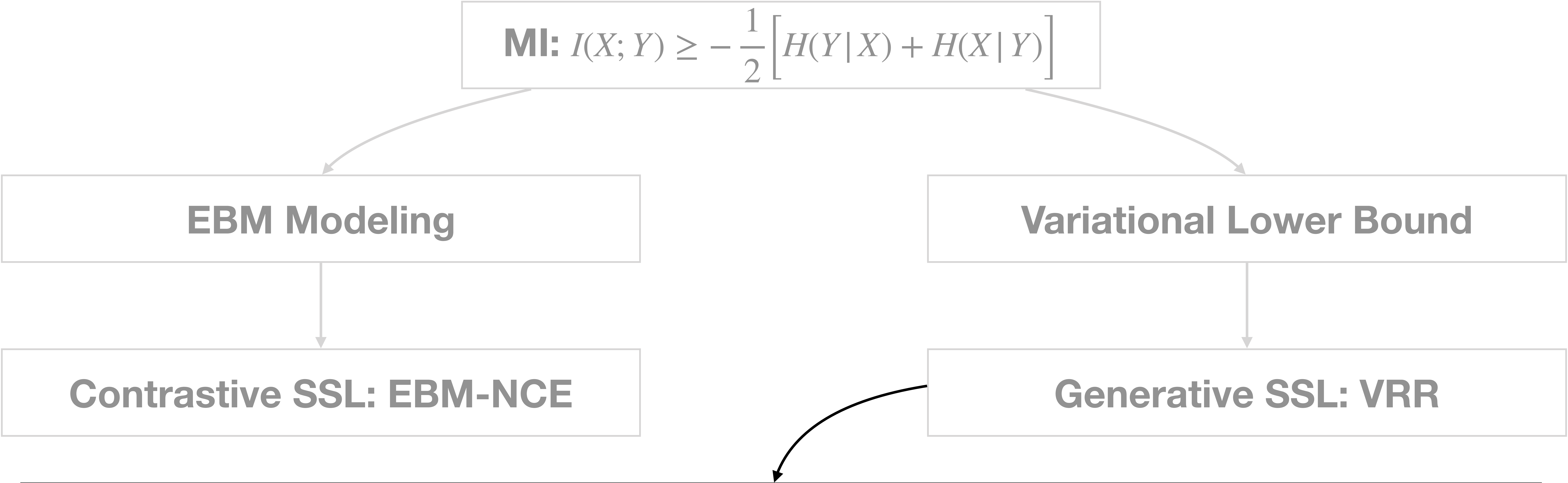
Contrastive SSL: EBM-NCE

Generative SSL: VRR

3. Another way to understand non-contrastive SSL.

- Q: If BYOL/SimSiam work, does this mean other generative SSL can also work well?

6 Future Directions: SSL on Structured Data



MI: $I(X;Y) \geq -\frac{1}{2}\Big[H(Y|X) + H(X|Y)\Big]$

EBM Modeling

Variational Lower Bound

Contrastive SSL: EBM-NCE

Generative SSL: VRR

3. Another way to understand non-contrastive SSL.

- Q: If BYOL/SimSiam work, does this mean other generative SSL can also work well?
- A: Yes! [1] provides the empirical evidence.

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

Thank you!
Q & A

Contrastive SSL

EBM-NCE & Jensen-Shannon Estimation (JSE)

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

- **Derivation and intuition:** Derivation process and underlying intuition are different.
    - JSE starts from f-divergence, then with variational estimation and Fenchel duality.
    - EBM-NCE is more straightforward: it models the conditional distribution in the MI lower bound with EBM, and solves it using NCE.
- **Noise distribution:** Starting from **MINE**, all the following works on graph SSL have been adopting the **empirical distribution** for **noise distribution**. However, this is not the case in EBM-NCE. Classic EBM-NCE uses fixed distribution, while more recent work extends it with adaptively learnable noise distribution.
- **Flexibility:** Modeling the conditional distribution with EBM provides a **broader family** of algorithms. **NCE** is just one solution to it, and recent progress on **score matching** and **contrastive divergence**, provides more promising directions.

EBM-NCE & InfoNCE

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

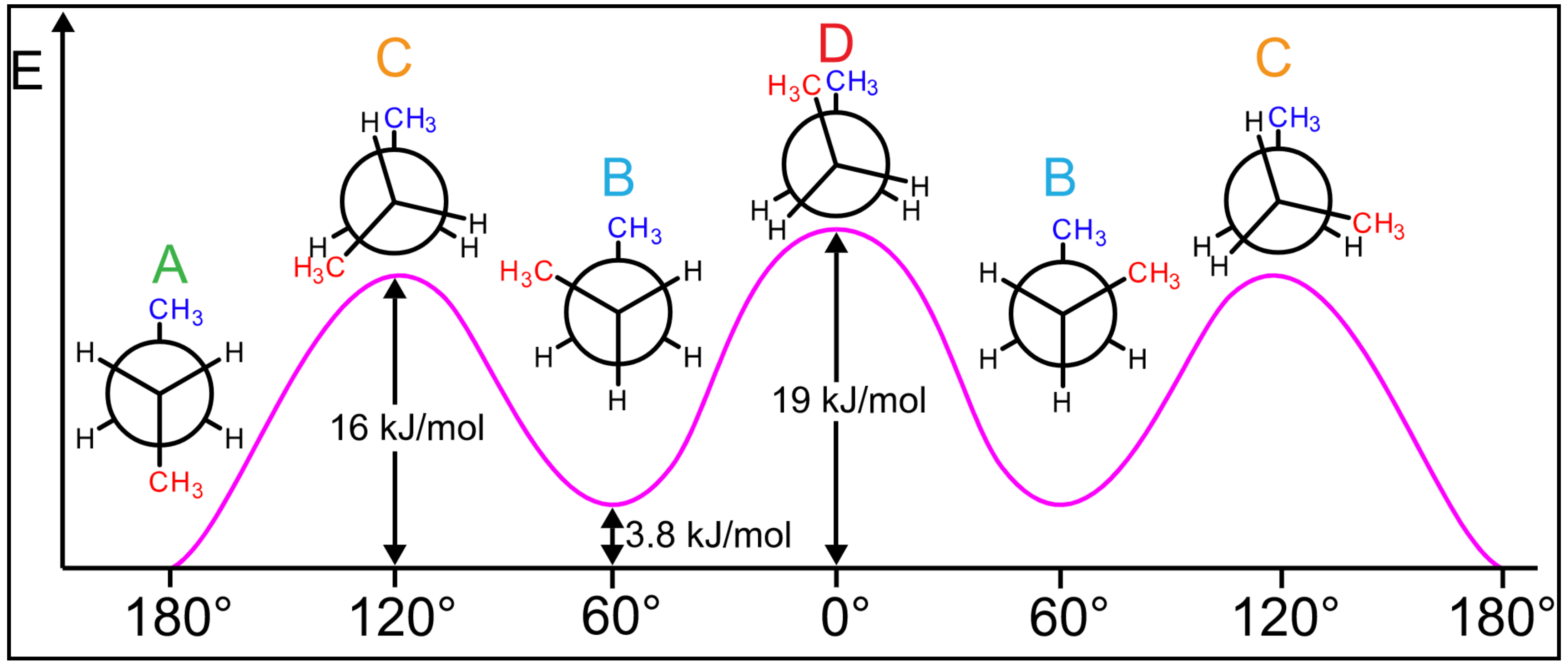
Empirically, EBM-NCE/JSE are more widely used in graph SSL.

Generative SSL

Lower bound on MI:

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

Credit to WiKi.



Variational Molecule Reconstruction

We introduce a variational distribution $z_x \sim \mathcal{N}(z_x; \mu_x, \Sigma_x)$:

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

Benefits:

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

Experiments

Ablation study on the impact of different objective functions.

- InfoNCE v.s. EBM-NCE
- VRR v.s. RR

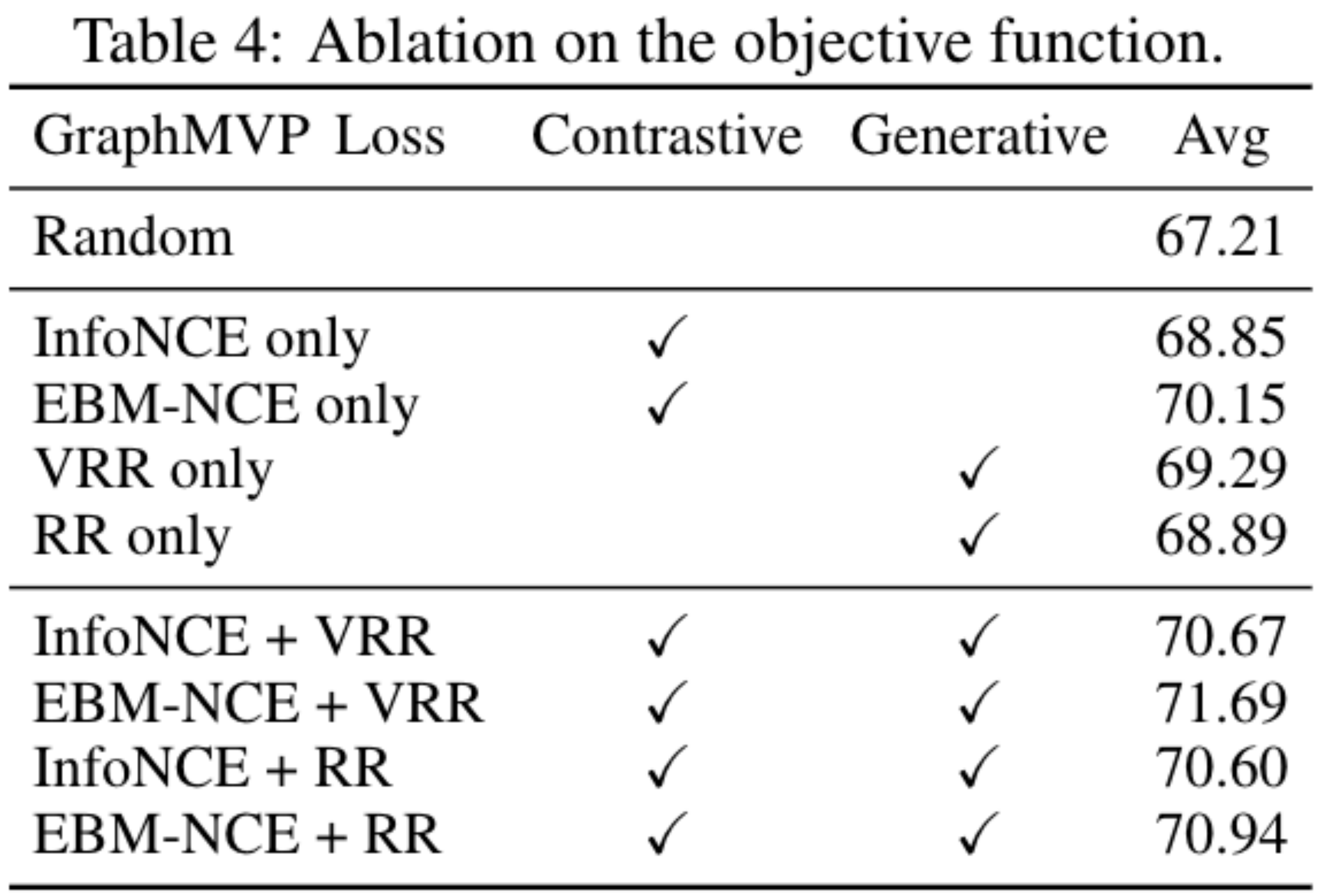


Table 4: Ablation on the objective function.

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

Experiments

Why SchNet?

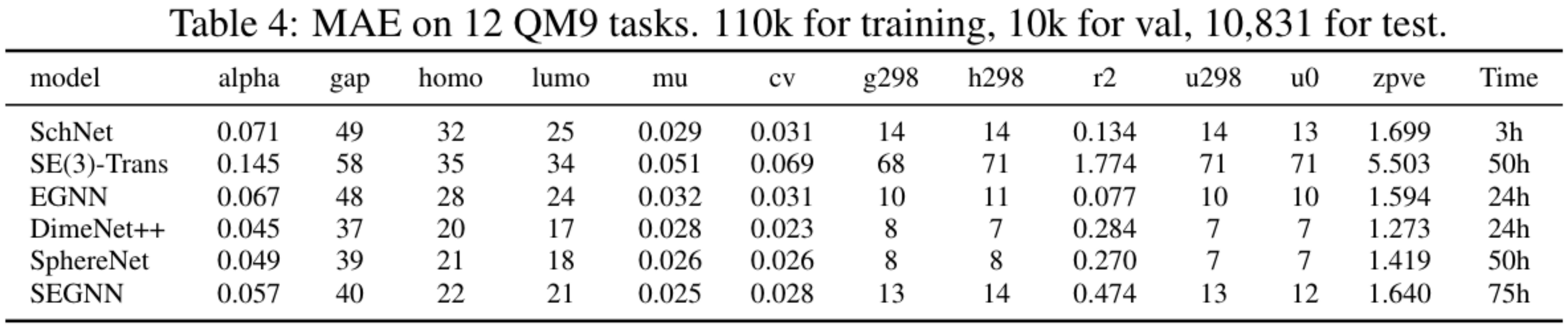


Table 4: MAE on 12 QM9 tasks. 110k for training, 10k for val, 10,831 for test.

model	alpha	gap	homo	lumo	mu	cv	g298	h298	r2	u298	u0	zpve	Time
SchNet	0.071	49	32	25	0.029	0.031	14	14	0.134	14	13	1.699	3h
SE(3)-Trans	0.145	58	35	34	0.051	0.069	68	71	1.774	71	71	5.503	50h
EGNN	0.067	48	28	24	0.032	0.031	10	11	0.077	10	10	1.594	24h
DimeNet++	0.045	37	20	17	0.028	0.023	8	7	0.284	7	7	1.273	24h
SphereNet	0.049	39	21	18	0.026	0.026	8	8	0.270	7	7	1.419	50h
SEGNN	0.057	40	22	21	0.025	0.028	13	14	0.474	13	12	1.640	75h