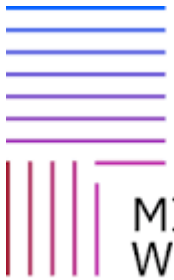
Improving Self-supervised Molecular Representation Learning using Persistent Homology

Yuankai Luo, Lei Shi, Veronika Thost

北京航空航天大學
BEIHANG UNIVERSITY



MIT-IBM
Watson
AI Lab

NeurIPS 2023

NEURAL INFORMATION
PROCESSING SYSTEMS

Motivation: SSL + Persistent Homology = ?

- Self-supervised learning (SSL) has great potential for molecular representation learning.

- Persistent homology (PH) is a mathematical tool for modeling topological features of data that persist across multiple scales.

- PH has proven effective for supervised molecular representation learning, esp. in studies from chemists.

- There are no studies on SSL!

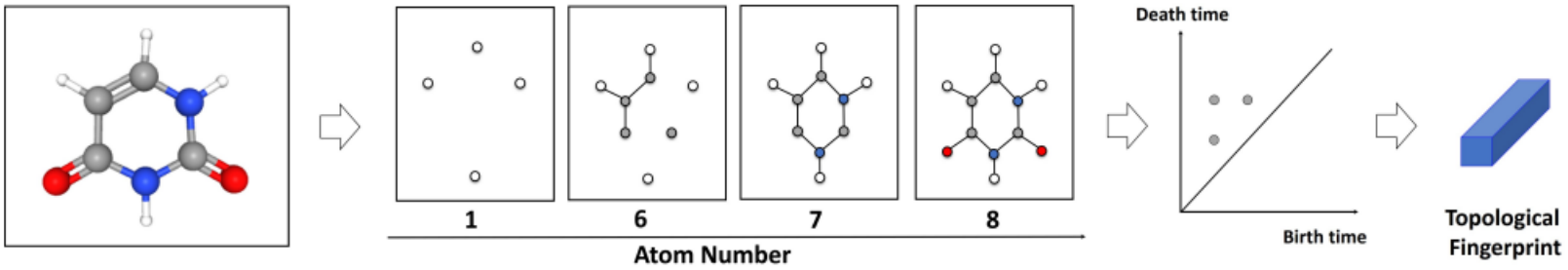
Persistent Homology (PH) on molecular graphs

- Molecules are graphs $G = (V, E)$ with nodes (0-simplex) V the atoms, and bond edges (1-simplex) E. Graph homology considers such a graph G as a topological space.

1. Filtrations. Construct a nested sequence of subgraphs $G_1 \subseteq \ldots \subseteq G_N = G$ by filtering, e.g., nodes by atom number.

2. Persistence Diagram (PD). During filtration, PH records all these birth and death times of the topological structures (the homology groups generated by simplices) in a PD.

3. Vectorization. Convert the PD into a format usable for ML called fingerprint, e.g., persistence images (PIs).



Persistent Homology (PH) on molecular graphs

Various opportunities for SSL

- Different filtrations and vectorizations yield views

- Stability feature of many fingerprints: distances between fingerprints are bounded by 1-WD between corresponding PDs

- Filtration design based on domain knowledge

We Explore the Potential of PH for SSL

Topological Fingerprints AutoEncoder (TAE)

- Here, we consider topological fingerprints I_G as the reconstruction targets:

$$h_G = R(g(\varepsilon(G)))$$

$$\mathcal{L}_{\mathrm{TAE}} = \sum_{\mathrm{G}} \mathrm{MSE}(h_G, I_G)$$

through a typical graph encoder $\varepsilon(G)$, a projection head $g(\cdot)$ and readout function $R(\cdot)$.

Topological Fingerprints AutoEncoder (TAE)

- Here, we consider topological fingerprints I_G as the reconstruction targets:

$$h_G = R(g(\varepsilon(G)))$$

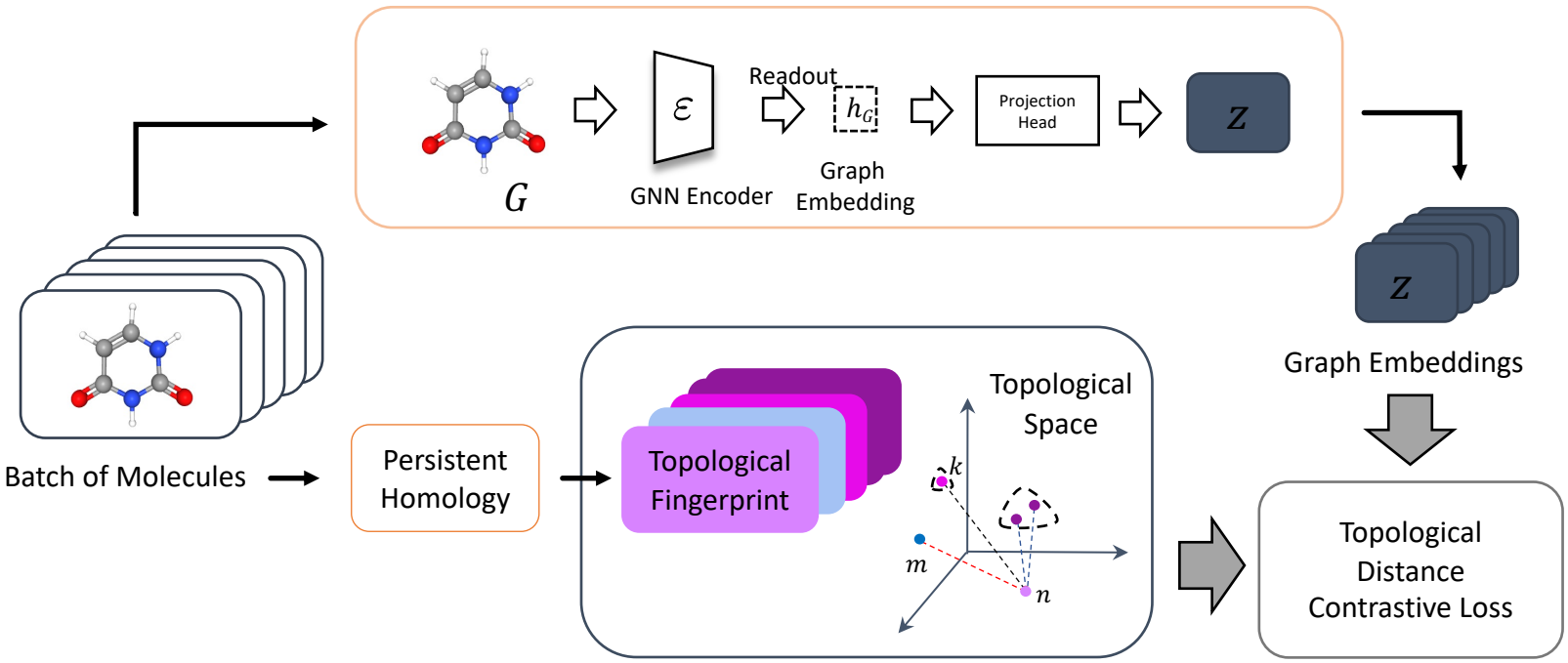
$$\mathcal{L}_{\mathrm{TAE}} = \sum_{\mathrm{G}} \mathrm{MSE}(h_G, I_G)$$

through a typical graph encoder $\varepsilon(G)$, a projection head $g(\cdot)$ and readout function $R(\cdot)$.

- Pre-trained TAE reconstructed downstream tasks' PIs (Pearson correlation coefficient)

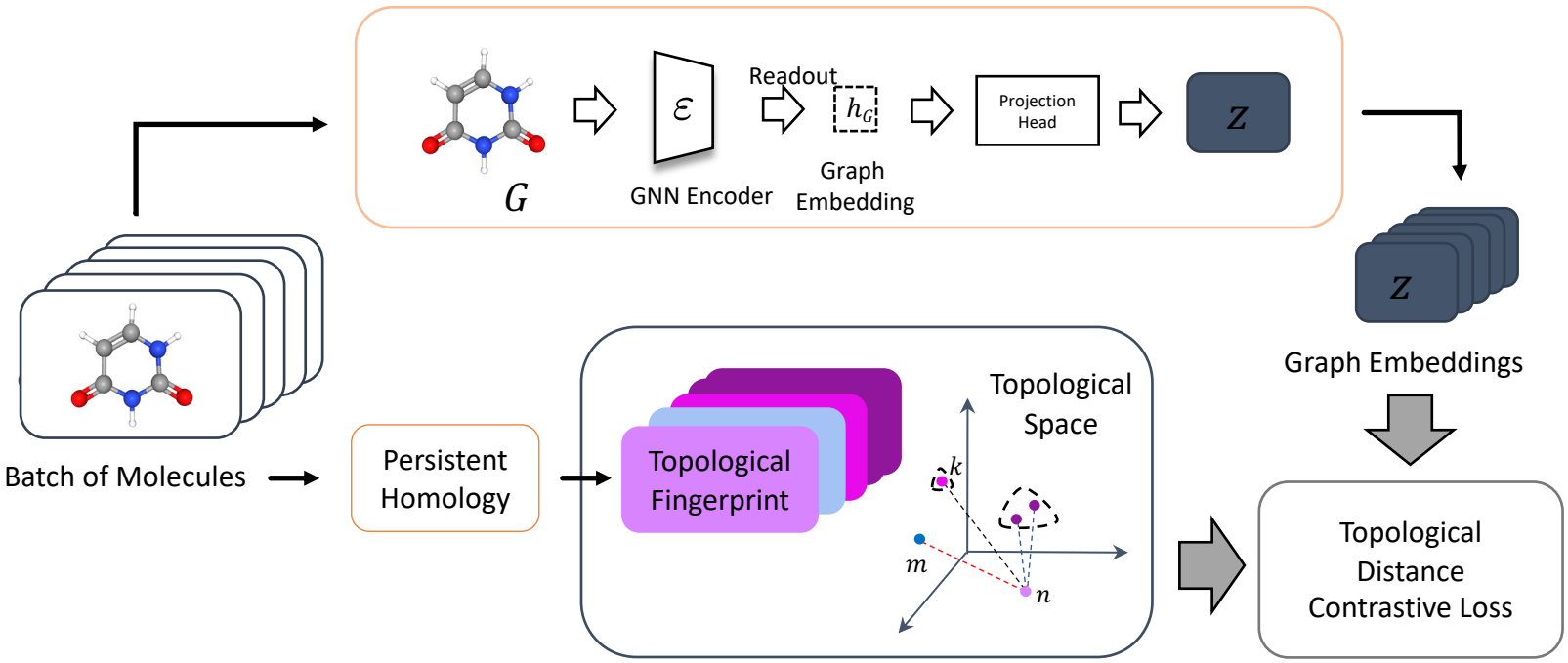
	Tox21	ToxCast	Sider	ClinTox	MUV	HIV	BBBP	Bace
# Molecules	7,831	8,575	1,427	1,478	93,087	41,127	2,039	1,513
# Molecules in ZINC15	628 (8%)	608 (7%)	1 (0%)	51 (4%)	7599 (8%)	925 (2%)	100 (5%)	0 (0%)
TAE	0.8572	0.7744	0.5939	0.8642	0.9044	0.7359	0.8660	0.8514

Topological Distance Contrastive Loss (TDL)



Different from regular contrastive learning, we have supervision about the distances between all molecules.

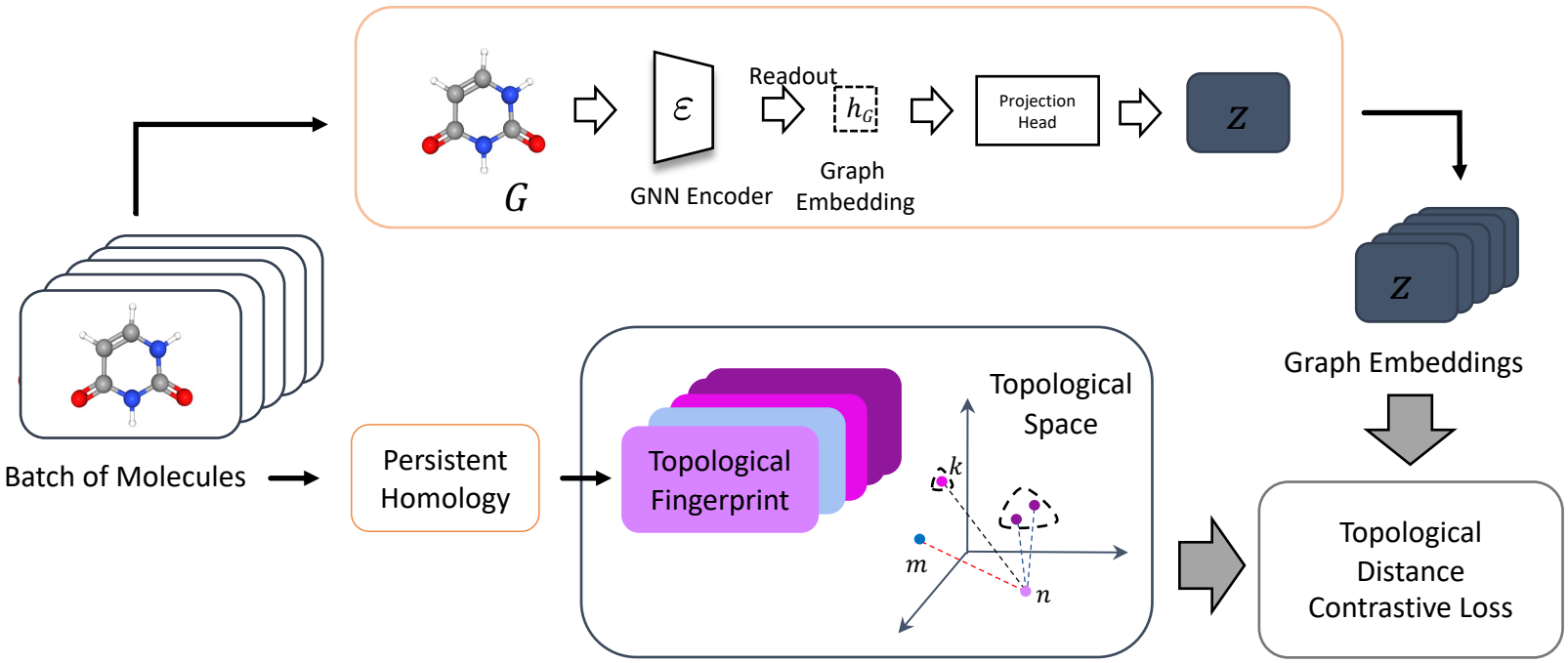
Topological Distance Contrastive Loss (TDL)



- Here, we focus on the distances between the given molecules (i.e., not views) since those are usually ignored.
- Our topological distance contrastive loss (TDL):

$$\mathcal{L}_{\mathrm{TDL}_n} = \frac{1}{N-1} \sum_{m \in [1,N], m \neq n} -\log \frac{\exp\left(sim\left(z_n, z_m\right)/\tau\right)}{\sum_{k \in [1,N], k \neq n} \mathbb{I}_{[dis(I_n, I_k) \geq dis(I_n, I_m)]} \cdot \exp\left(sim\left(z_n, z_k\right)/\tau\right)}$$

Topological Distance Contrastive Loss (TDL)



TDL is efficient and can be flexibly applied to improve the embedding space (the main goal of SSL) of any existing contrastive method.

Evaluation

Table 4: Binary classification over MoleculeNet; ROC-AUC, % Pos. is min/med/max for multi-task.

	Tox21	ToxCast	Sider	ClinTox	MUV	HIV	BBBP	Bace	Average
# Molecules	7,831	8,575	1,427	1,478	93,087	41,127	2,039	1,513	-
# Tasks	12	617	27	2	17	1	1	1	
% Positives	2.4/4.6/12.0	0.2/1.3/20.5	1.5/66.3/92.4	7.6/50.6/93.6	0.03/0.03/0.03	3.5	76.5	45.7	
No pretrain (GIN)	74.6 (0.4)	61.7 (0.5)	58.2 (1.7)	58.4 (6.4)	70.7 (1.8)	75.5 (0.8)	65.7 (3.3)	72.4 (3.8)	67.15
AD-GCL [Suresh et al., 2021]	76.5 (0.8)	63.0 (0.7)	63.2 (0.7)	79.7 (3.5)	72.3 (1.6)	78.2 (0.9)	70.0 (1.0)	78.5 (0.8)	72.67
iMolCLR [Wang et al., 2022b]	75.1 (0.7)	63.5 (0.4)	59.4 (1.0)	81.0 (2.6)	74.7 (1.9)	77.3 (1.2)	69.6 (1.2)	77.3 (1.0)	72.24
Mole-BERT [Xia et al., 2023b]	76.8 (0.5)	64.3 (0.2)	62.8 (1.1)	78.9 (3.0)	78.6 (1.8)	78.2 (0.8)	71.9 (1.6)	80.8 (1.4)	74.04
SEGA [Wu et al., 2023]	76.7 (0.4)	65.2 (0.9)	63.6 (0.3)	84.9 (0.9)	76.6 (2.4)	77.6 (1.3)	71.8 (1.0)	77.0 (0.4)	74.17
TAE_{ahd}	75.2 (0.8)	63.1 (0.3)	61.9 (0.8)	80.6 (1.9)	74.6 (1.8)	73.5 (2.1)	67.5 (1.1)	82.5 (1.1)	72.36
TAE_{ToDD}	76.8 (0.9)	64.0 (0.5)	61.9 (0.8)	79.3 (3.6)	75.8 (3.2)	75.9 (1.1)	70.4 (0.8)	81.6 (1.4)	73.22
ContextPred	75.7 (0.7)	63.9 (0.6)	60.9 (0.6)	65.9 (3.8)	75.8 (1.7)	77.3 (1.0)	68.0 (2.0)	79.6 (1.2)	70.89
+ TAE_{ahd}	76.4 (0.5)	63.2 (0.4)	62.0 (0.7)	74.6 (4.4)	76.7 (1.6)	77.7 (1.2)	68.9 (1.1)	80.7 (1.6)	**72.53**
+ TAE_{ToDD}	75.7 (0.4)	63.1 (0.3)	61.3 (0.5)	72.1 (1.3)	77.2 (1.8)	77.6 (1.1)	69.6 (0.9)	80.1 (1.4)	**72.09**
GraphCL	73.9 (0.7)	62.4 (0.6)	60.5 (0.9)	76.0 (2.7)	69.8 (2.7)	78.5 (1.2)	69.7 (0.7)	75.4 (1.4)	70.78
+ TDL_{atom}	75.3 (0.4)	64.4 (0.3)	61.2 (0.6)	83.7 (2.7)	75.7 (0.8)	78.0 (0.9)	70.9 (0.6)	80.5 (0.8)	**73.71**
+ TDL_{ToDD}	75.2 (0.7)	64.2 (0.3)	61.5 (0.4)	85.2 (1.8)	75.9 (2.1)	77.9 (0.8)	69.9 (0.9)	81.2 (1.9)	**73.88**
JOAO	75.0 (0.3)	62.9 (0.5)	60.0 (0.8)	81.3 (2.5)	71.7 (1.4)	76.7 (1.2)	70.2 (1.0)	77.3 (0.5)	71.89
+ TDL_{atom}	75.5 (0.3)	63.8 (0.2)	60.6 (0.5)	76.8 (1.5)	73.8 (1.9)	78.3 (1.2)	70.3 (0.5)	78.7 (0.6)	**72.22**
+ TDL_{ToDD}	75.2 (0.3)	63.6 (0.2)	61.6 (0.6)	80.7 (3.3)	74.6 (1.6)	77.4 (0.9)	71.3 (0.8)	81.0 (2.2)	**73.18**
SimGRACE	74.4 (0.3)	62.6 (0.7)	60.2 (0.9)	75.5 (2.0)	75.4 (1.3)	75.0 (0.6)	71.2 (1.1)	74.9 (2.0)	71.15
+ TDL_{atom}	74.7 (0.5)	63.0 (0.3)	59.5 (0.4)	73.7 (1.5)	75.9 (1.6)	77.3 (1.1)	69.5 (0.9)	79.1 (0.5)	**71.59**
+ TDL_{ToDD}	75.6 (0.4)	63.3 (0.5)	59.9 (0.8)	82.4 (2.5)	75.6 (2.0)	76.1 (1.3)	69.9 (0.8)	78.9 (1.6)	**72.71**
GraphLoG	75.0 (0.6)	63.4 (0.6)	59.3 (0.8)	70.1 (4.6)	75.5 (1.6)	76.1 (0.8)	69.6 (1.6)	82.1 (1.0)	71.43
+ TDL_{atom}	76.1 (0.7)	63.7 (0.4)	59.9 (1.0)	75.7 (3.5)	75.7 (1.2)	76.2 (1.8)	69.6 (1.2)	82.2 (1.5)	**72.39**
+ TDL_{ToDD}	75.9 (0.8)	63.5 (0.7)	63.4 (0.3)	79.8 (1.9)	75.6 (1.1)	76.2 (1.6)	70.7 (0.9)	82.1 (1.9)	**73.39**

- Notably, TDL demonstrates convincing improvements across all baselines and gets competitive with SOTA.

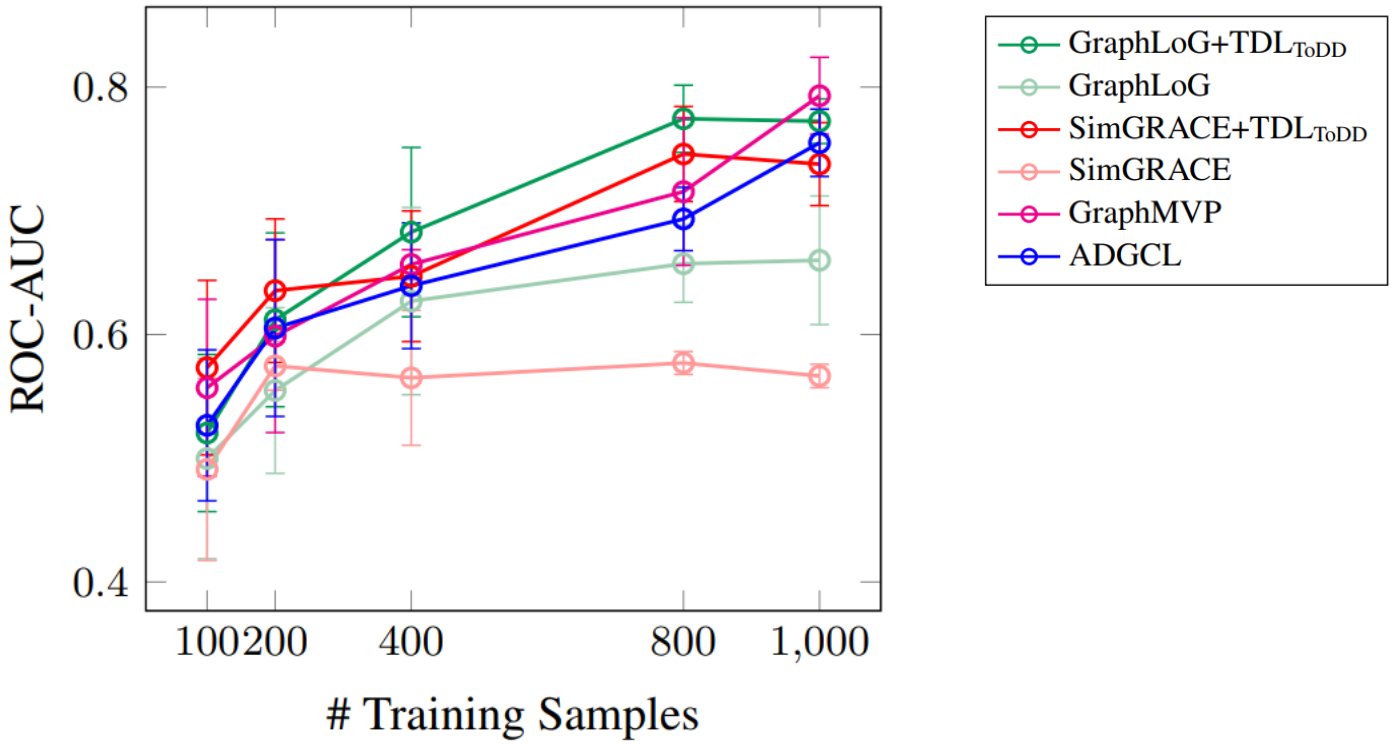
Evaluation

Table 2: Linear/MLP probing: molecular property prediction; binary classification, ROC-AUC (%).

ECFP, MLP	70.1 (0.4)	59.8 (0.4)	59.6 (0.6)	67.8 (0.9)	61.7 (0.8)	69.1 (1.0)	58.6 (1.3)	72.1 (1.7)	64.85
ECFP $\|\|$ PI_{ToDD}, MLP	71.1 (0.6)	57.8 (0.4)	59.2 (0.7)	80.7 (2.1)	64.9 (1.1)	72.8 (1.7)	63.1 (0.8)	76.7 (0.9)	**68.28**
TAE_{ahd}	67.7 (0.2)	61.2 (0.2)	55.8 (0.3)	58.1 (0.7)	70.2 (0.8)	72.5 (0.5)	61.1 (0.2)	74.3 (0.2)	65.11
TAE_{ToDD}	70.4 (0.2)	60.8 (0.1)	61.1 (0.1)	68.4 (0.7)	72.3 (0.3)	73.9 (0.2)	61.6 (0.4)	67.6 (0.6)	67.01
ContextPred	68.4 (0.3)	59.1 (0.2)	59.4 (0.3)	43.2 (1.7)	71.0 (0.7)	68.9 (0.4)	59.1 (0.2)	64.4 (0.6)	61.69
+ TAE_{ahd}	69.7 (0.1)	59.2 (0.2)	59.5 (0.3)	56.1 (1.1)	76.5 (0.9)	68.9 (0.2)	61.1 (0.4)	65.6 (0.5)	**64.58**
+ TAE_{ToDD}	69.0 (0.1)	59.8 (0.4)	60.0 (0.4)	53.3 (1.3)	70.8 (0.3)	70.0 (0.7)	60.9 (0.5)	62.7 (0.5)	**63.31**
GraphCL	64.4 (0.5)	59.4 (0.2)	54.6 (0.3)	59.8 (1.2)	70.2 (1.0)	63.7 (2.3)	62.4 (0.7)	71.1 (0.7)	63.20
+ TDL_{atom}	72.0 (0.4)	61.1 (0.2)	59.7 (0.6)	65.3 (1.3)	76.1 (0.9)	68.2 (1.1)	65.4 (0.9)	76.4 (1.1)	**68.02**
+ TDL_{ToDD}	72.7 (0.5)	60.8 (0.4)	58.9 (0.8)	64.1 (1.7)	72.7 (1.4)	69.7 (1.2)	64.5 (0.8)	76.1 (1.3)	**67.44**
JOAO	70.6 (0.4)	60.5 (0.3)	57.4 (0.6)	54.1 (2.6)	69.8 (1.9)	68.1 (0.9)	63.7 (0.3)	71.2 (1.0)	64.42
+ TDL_{atom}	70.5 (0.3)	60.4 (0.2)	57.8 (1.5)	54.6 (1.3)	74.2 (1.6)	68.2 (0.6)	65.2 (0.3)	72.7 (3.1)	**65.41**
+ TDL_{ToDD}	71.7 (0.4)	61.3 (0.3)	58.9 (0.7)	52.4 (1.7)	69.6 (1.7)	69.9 (0.6)	64.1 (0.5)	72.6 (0.9)	**65.06**
SimGRACE	64.6 (0.4)	59.1 (0.2)	54.9 (0.6)	63.4 (2.6)	67.4 (1.2)	66.3 (1.5)	65.4 (1.2)	67.8 (1.3)	63.61
+ TDL_{atom}	68.6 (0.3)	61.1 (0.2)	59.5 (0.4)	62.2 (1.7)	69.7 (2.0)	69.5 (1.8)	60.6 (0.5)	72.1 (0.7)	**65.41**
+ TDL_{ToDD}	70.1 (0.3)	60.3 (0.3)	59.1 (0.3)	65.1 (1.4)	71.4 (1.1)	71.1 (0.7)	64.9 (0.6)	73.4 (0.8)	**66.93**
GraphLoG	67.2 (0.2)	57.9 (0.2)	57.9 (0.3)	57.8 (0.9)	64.2 (1.1)	65.0 (1.3)	54.3 (0.7)	72.3 (0.9)	62.08
+ TDL_{atom}	72.1 (0.3)	62.0 (0.2)	60.7 (0.2)	56.6 (0.8)	73.0 (0.9)	70.4 (0.9)	61.2 (0.4)	76.8 (0.7)	**66.59**
+ TDL_{ToDD}	70.7 (0.2)	60.7 (0.3)	61.5 (0.3)	59.5 (0.5)	72.9 (1.8)	71.6 (0.8)	62.1 (0.3)	80.1 (0.4)	**67.39**

- The results are mixed, TDL yields overall impressive increases.

Evaluation


ROC-AUC
0.8
0.6
0.4
100
200
400
800
1,000
Training Samples
GraphLoG+TDL$_{\text{ToDD}}$
GraphLoG
SimGRACE+TDL$_{\text{ToDD}}$
SimGRACE
GraphMVP
ADGCL

Conclusions

TDL is overall effective

- Particularly in probing and w/ low data, where the SSL embedding space is important.
- It also helps mitigating deficiencies of individual baselines.

https://github.com/LUOyk1999/Molecular-homology

Conclusions

TDL is overall effective

- Particularly in probing and w/ low data, where the SSL embedding space is important.
- It also helps mitigating deficiencies of individual baselines.

Thanks for listening!

https://github.com/LUOyk1999/Molecular-homology