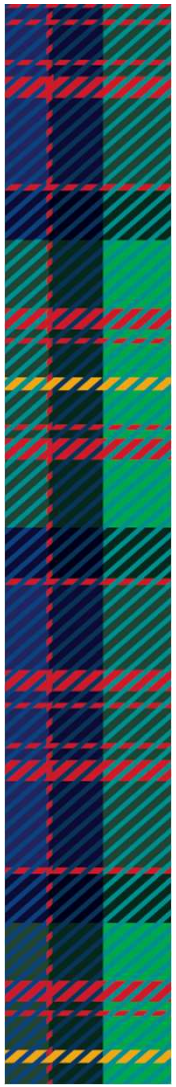
Contrastive Learning and Implementation on Molecules

Yuyang Wang, PhD Candidate
Department of Mechanical Engineering
Carnegie Mellon University
Email: yuyangw@cmu.edu

May 2021

Agenda

- ❑ Motivation
- ❑ Method
  - Contrastive Learning
  - Molecule Graph Augmentation
  - MolCLR
- ❑ Experimental Results
  - Classification
  - Regression
  - Visualization



Motivation

Motivation

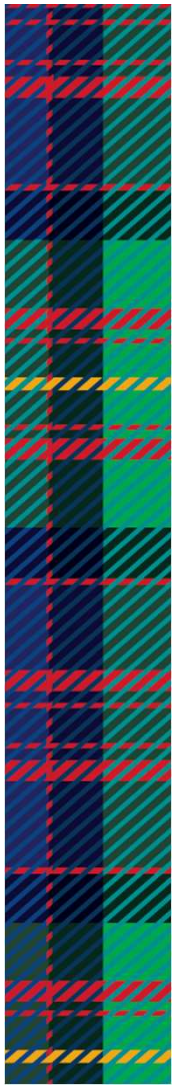
1. Hard to represent the molecular information thoroughly.

E.g., SMILES, SMART fail to encode the topology information directly

2. Enormous magnitude of chemical

E.g., the size of potential pharmacologically active molecules is estimated to be in the order of 10^{60}

3. Expensive and insufficient labeled data for molecular learning.



Method: MolCLR

Contrastive Learning

- Contrastive Loss is a Self-Supervised Learning strategy which aims at learning representation through ***contrasting positive data pairs against negative data pairs***. Normally, we refer to similar data as positive data pairs, and dissimilar data as negative pairs.
- The goal is to ***learn a representation*** which can be transferred easily to various downstream tasks, including classification and regr

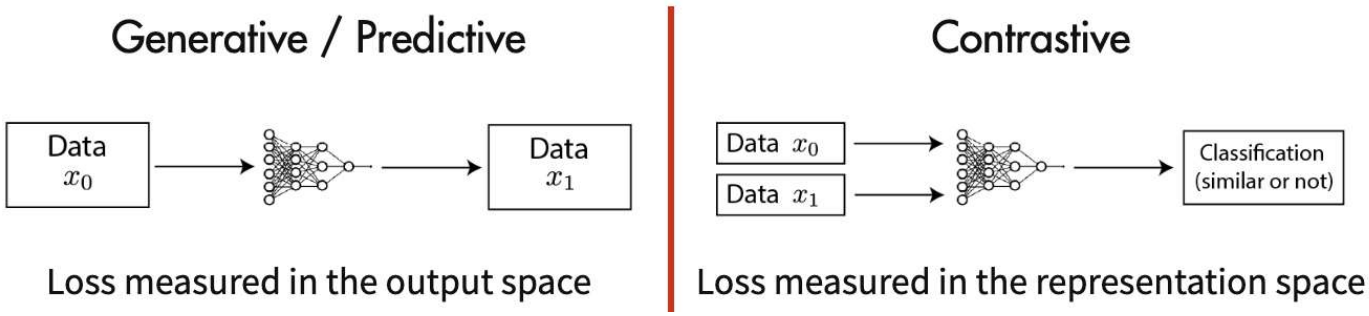


Figure 4: Contrastive Learning v.s. Generative/Predictive Learning

Molecule Graph Augmentation

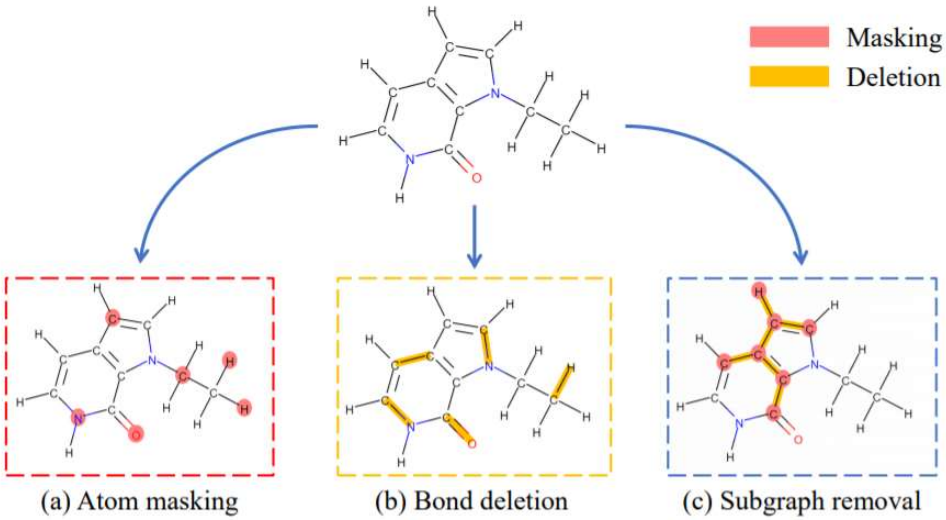


Figure 5: Three molecule graph augmentation strategies. (a) Atom masking: randomly replaces the node feature xv of an atom feature with a mask token m.(b) Bond deletion: randomly deletes the bond between two atoms, so that the they are not directly connected on the graph. (c) Subgraph removal: randomly removes an induced subgraph from the original molecule graph. Within the subgraph, all nodes are masked, and all edges are deleted.

MolCLR

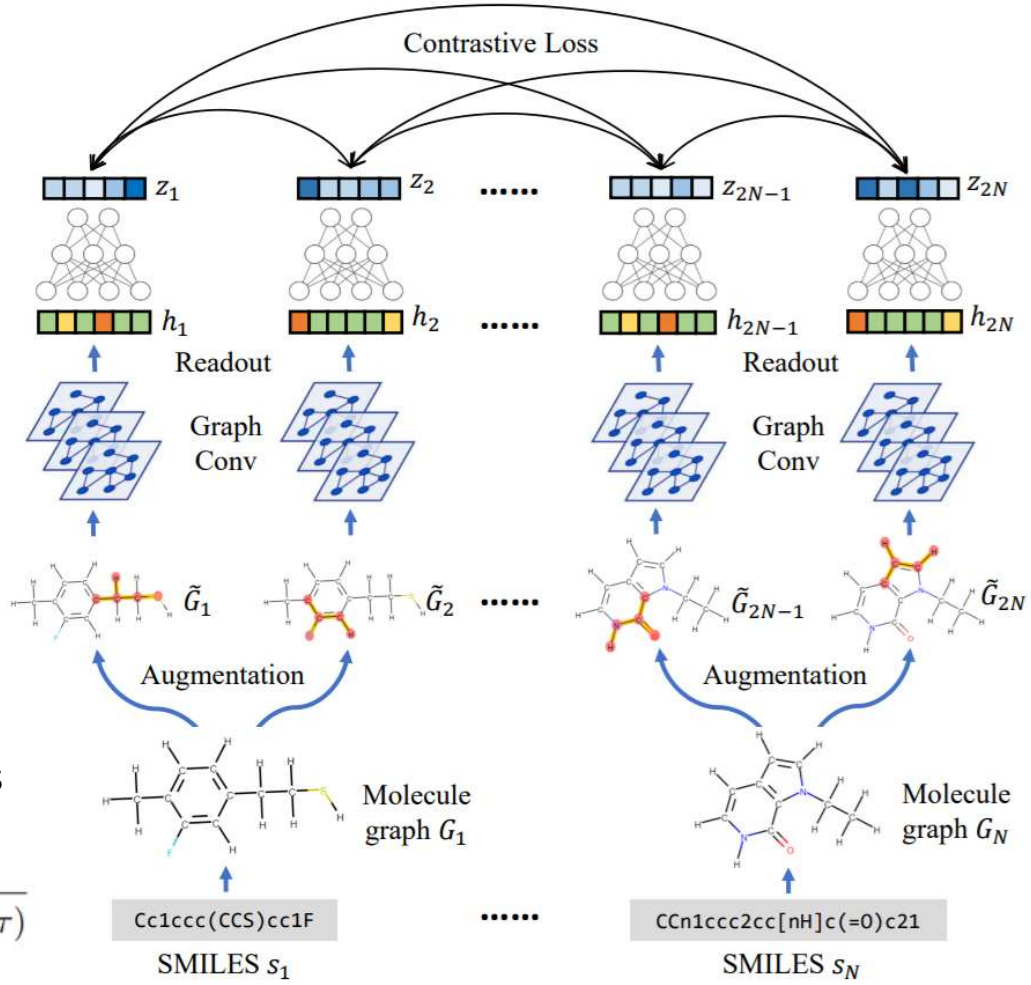


Figure 6: Pipeline of Molecular Contrastive Learning of Representations (MolCLR) via GNNs

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Experimental Results

Classification

Table 1: Test ROC-AUC (%) performance comparison of different models, where the first five models are supervised learning methods and the last three are self-supervised/pre-training methods. Mean and standard deviation on each benchmark are reported.

Dataset	BBBP	Tox21	ClinTox	HIV	BACE	SIDER	MUV
# Molecules	2039	7831	1478	41127	1513	1478	93087
# Tasks	1	12	2	1	1	27	17
RF	71.4±0.0	76.9±1.5	71.3±5.6	78.1±0.6	**86.7±0.8**	**68.4±0.9**	63.2±2.3
SVM	72.9±0.0	**81.8±1.0**	66.9±9.2	**79.2±0.0**	86.2±0.0	**68.2±1.3**	67.3±1.3
MGCN	**85.0±6.4**	70.7±1.6	63.4±4.2	73.8±1.6	73.4±3.0	55.2±1.8	70.2±3.4
D-MPNN	71.2±3.8	68.9±1.3	**90.5±5.3**	75.0±2.1	85.3±5.3	63.2±2.3	**76.2±2.8**
HU. et.al	70.8±1.5	78.7±0.4	78.9±2.4	80.2±0.9	85.9±0.8	65.2±0.9	81.4±2.0
N-Gram	**91.2±3.0**	76.9±2.7	85.5±3.7	**83.0±1.3**	87.6±3.5	63.2±0.5	81.6±1.9
MolCLR	73.6±0.5	**79.8±0.7**	**93.2±1.7**	80.6±1.1	**89.0±0.3**	**68.0±1.1**	**88.6±2.2**

Regression

Table 2: Test performance comparison of different models on regression tasks, where the first five models are supervised learning methods and the last three are self-supervised/pre-training methods. Mean and standard deviation on each benchmark are reported.

Dataset	FreeSolv	ESOL	Lipo	QM7	QM8
# Molecules	642	1128	4200	6830	21786
# Tasks	1	1	1	1	12
RF	**2.03±0.22**	1.07±0.19	0.88±0.04	122.7±4.2	0.042±0.002
GCN [1]	2.900±0.135	1.068±0.050	0.712±0.049	118.9±20.2	0.021±0.001
SchNet [2]	3.22±0.76	1.05±0.06	0.91±0.10	**74.2±6.0**	0.020±0.002
MGCN [3]	3.35±0.01	1.27±0.15	1.11±0.04	77.6±4.7	0.022±0.002
D-MPNN [4]	2.18±0.91	**0.98±0.26**	**0.65±0.05**	105.8±13.2	**0.0143±0.002**
N-Gram [5]	2.51±0.19	**1.10±0.03**	0.88±0.12	125.6±1.5	0.0320±0.003
MolCLR	**2.20±0.20**	**1.11±0.01**	**0.65±0.08**	**87.2±2.0**	**0.0174±0.001**

MolCLR Visualization

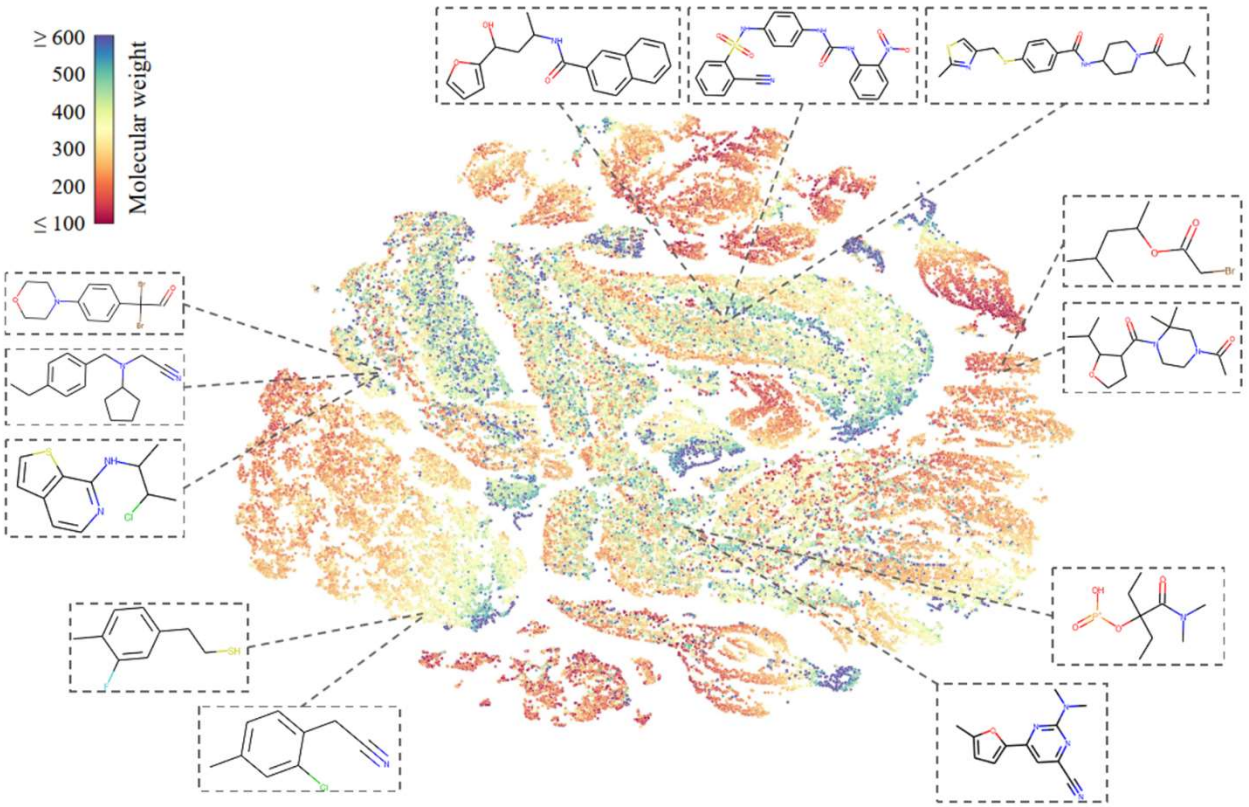


Figure 7: Two-dimensional t-SNE embedding of the molecular representations learned by MolCLR pre-training. The color of each embedding point indicates its corresponding molecular weight.

MolCLR Visualization

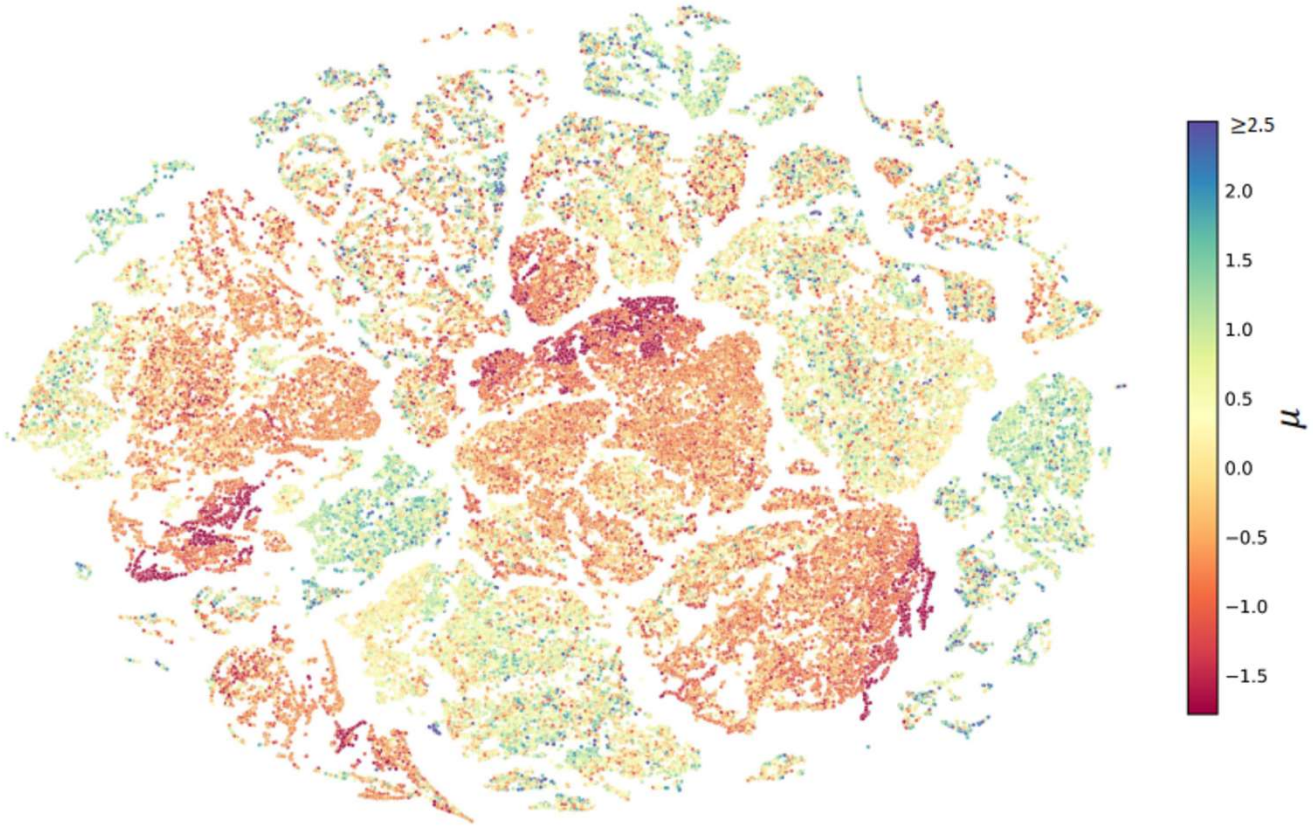


Figure 8: Two-dimensional t-SNE embedding of the molecular representations learned by MolCLR pre-training. The color of each embedding point indicates the averaged electronic spectrum of each molecule.

Thanks!
Q&A