Molecular Contrastive Learning of Representations via Graph Neural Networks

School of industrial and Management Engineering, Korea University

Young Jae Lee

KOREA UNIVERSITY

DMQA

Contents

Research Purpose

❖ Molecular Contrastive Learning of Representations via Graph Neural Networks (arXiv, 2021)

- Carnegie Mellon 대학에서 연구하였고 2022년 03월 03일 기준 약 17회 인용

Molecular Contrastive Learning of Representations via Graph Neural Networks

Yuyang Wang[1,2], Jianren Wang[3], Zhonglin Cao[1], and Amir Barati Farimani[1,2,4,*]

[1]Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[4]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[*]corresponding author: Amir Barati Farimani (barati@cmu.edu)

ABSTRACT

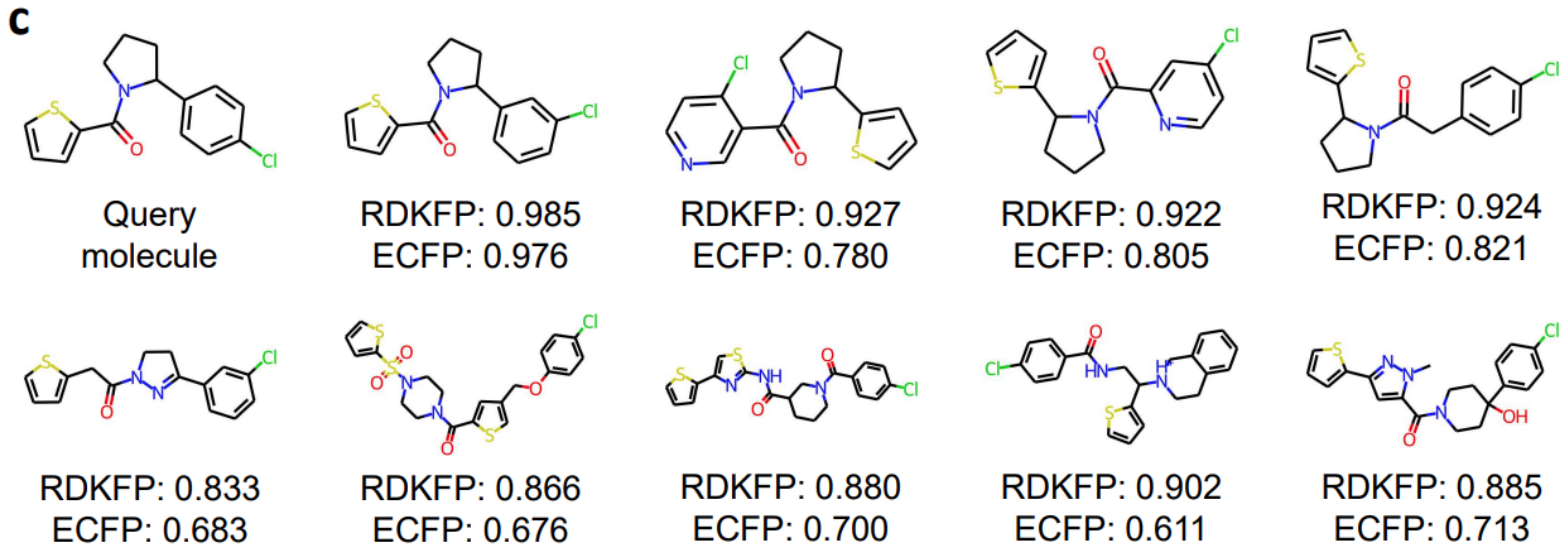
Molecular Machine Learning (ML) bears promise for efficient molecule property prediction and drug discovery. However, labeled molecule data can be expensive and time-consuming to acquire. Due to the limited labeled data, it is a great challenge for supervised-learning ML models to generalize to the giant chemical space. In this work, we present *MolCLR*: Molecular Contrastive Learning of Representations via Graph Neural Networks (GNNs), a self-supervised learning framework that leverages large unlabeled data (~10M unique molecules). In MolCLR pre-training, we build molecule graphs and develop GNN encoders to learn differentiable representations. Three molecule graph augmentations are proposed: atom masking, bond deletion, and subgraph removal. A contrastive estimator maximizes the agreement of augmentations from the same molecule while minimizing the agreement of different molecules. Experiments show that our contrastive learning framework significantly improves the performance of GNNs on various molecular property benchmarks including both classification and regression tasks. Benefiting from pre-training on the large unlabeled database, *MolCLR* even achieves state-of-the-art on several challenging benchmarks after fine-tuning. Additionally, further investigations demonstrate that *MolCLR* learns to embed molecules into representations that can distinguish chemically reasonable molecular similarities.

Research Purpose

Graph

❖ Molecular Contrastive Learning of Representations via Graph Neural Networks (arXiv, 2021)

- 화학 분야 중 지도 학습 기반 분자 성질 예측, 신약 발견 분야에서 큰 성공을 거두었음
- 하지만, 레이블이 있는 데이터를 수집하는데 **비용, 시간 소비가 매우 큼**
- 레이블이 없는 데이터를 사용하여 **분자 정보를 세밀하게 표현하는 것**이 핵심
- 본 연구에서는 천만 개의 레이블이 없는 데이터(PubChem)를 사용한 대조 학습 방법을 제안
  - ✓ 특히, 분자 구조를 그래프로 표현하여 **그래프 기반 대조 학습** 제안



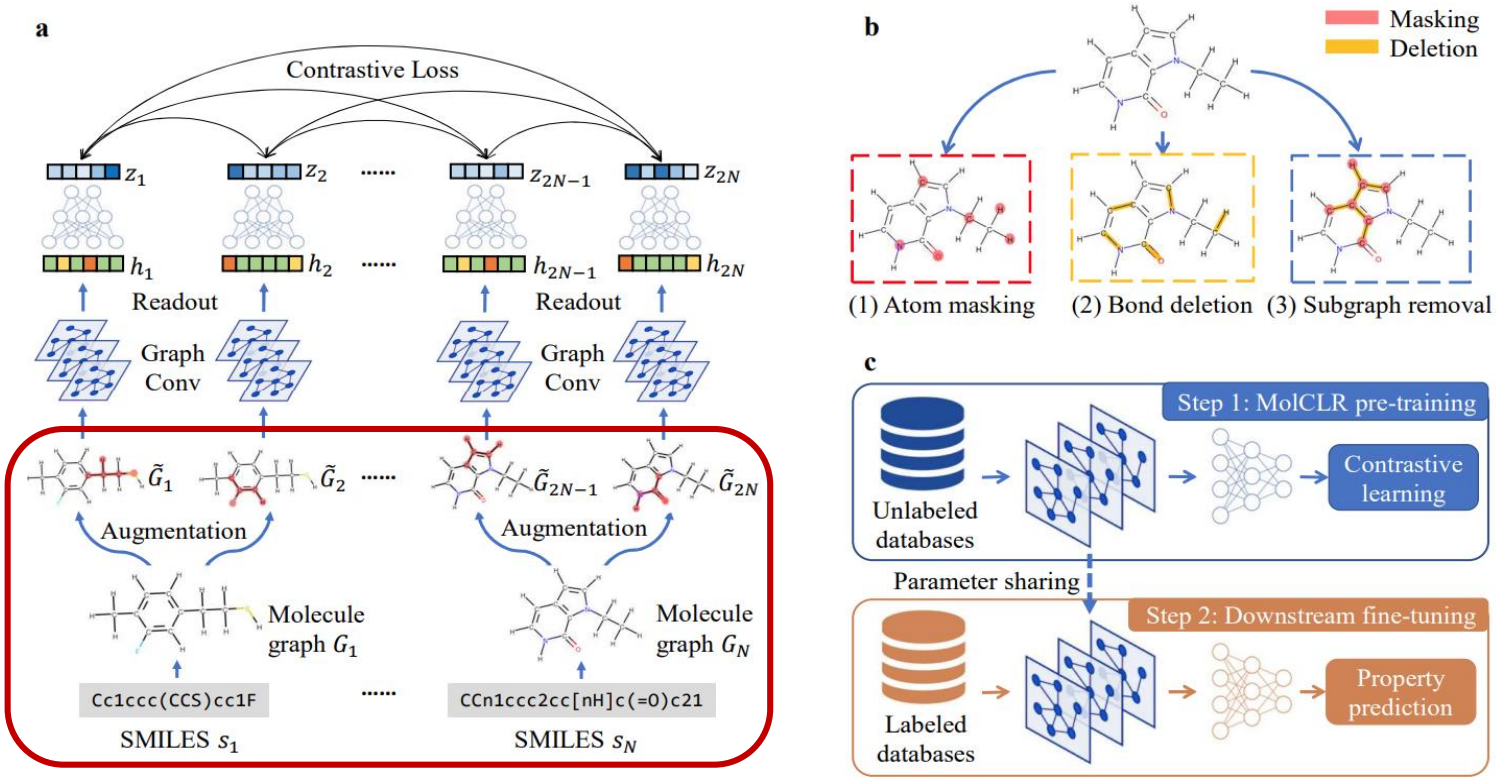
DMQA

MolCLR

Graph

❖ **Mol**ecular **C**ontrastive **L**earning of **R**epresentations (MolCLR)

- 분자 그래프 표현 학습을 위해 SimCLR 적용 – Positive Pair와 Negative Examples 정의
  - ✓ 시퀀스 형태로 표현되는 분자(SMILES String)를 그래프 형태로 변환
  - ✓ 한 분자 그래프에서 서로 다른 증강 기법을 적용한 쌍 Positive Pair $(\tilde{G}_1, \tilde{G}_2)$, 이 외에는 Negative Examples



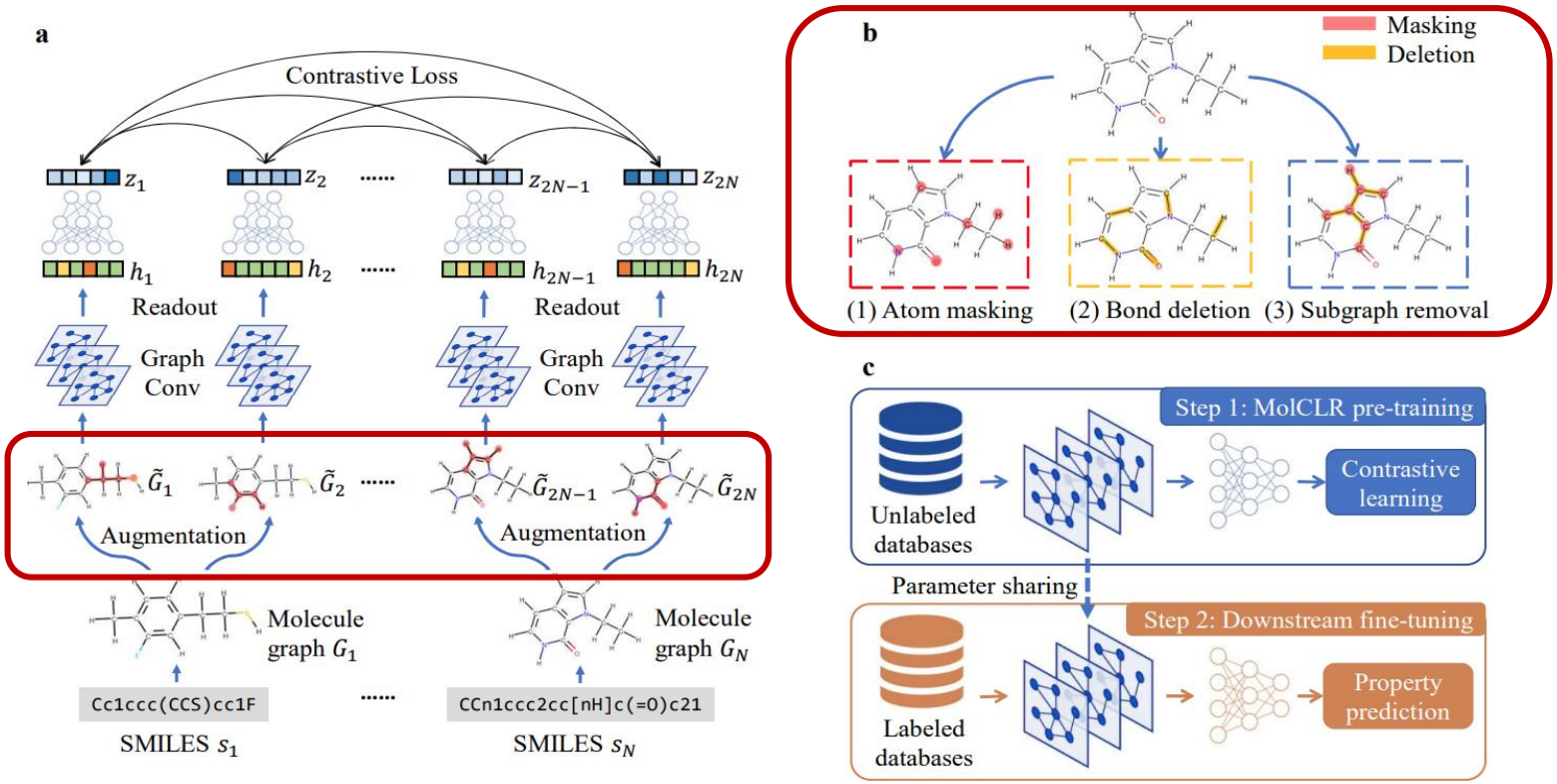
DMQA

MolCLR

Graph

❖ **Mol**ecular **C**ontrastive **L**earning of **R**epresentations (MolCLR)

- 분자 그래프 표현 학습을 위해 SimCLR 적용 – 데이터 증강 기법
  - ✓ Atom Masking, Bond Deletion, Subgraph Removal 적용
  - ✓ 일정 비율로 원자 정보를 삭제, 원자 간 결합 정보를 삭제, 원자와 결합 정보 일부를 통으로 삭제



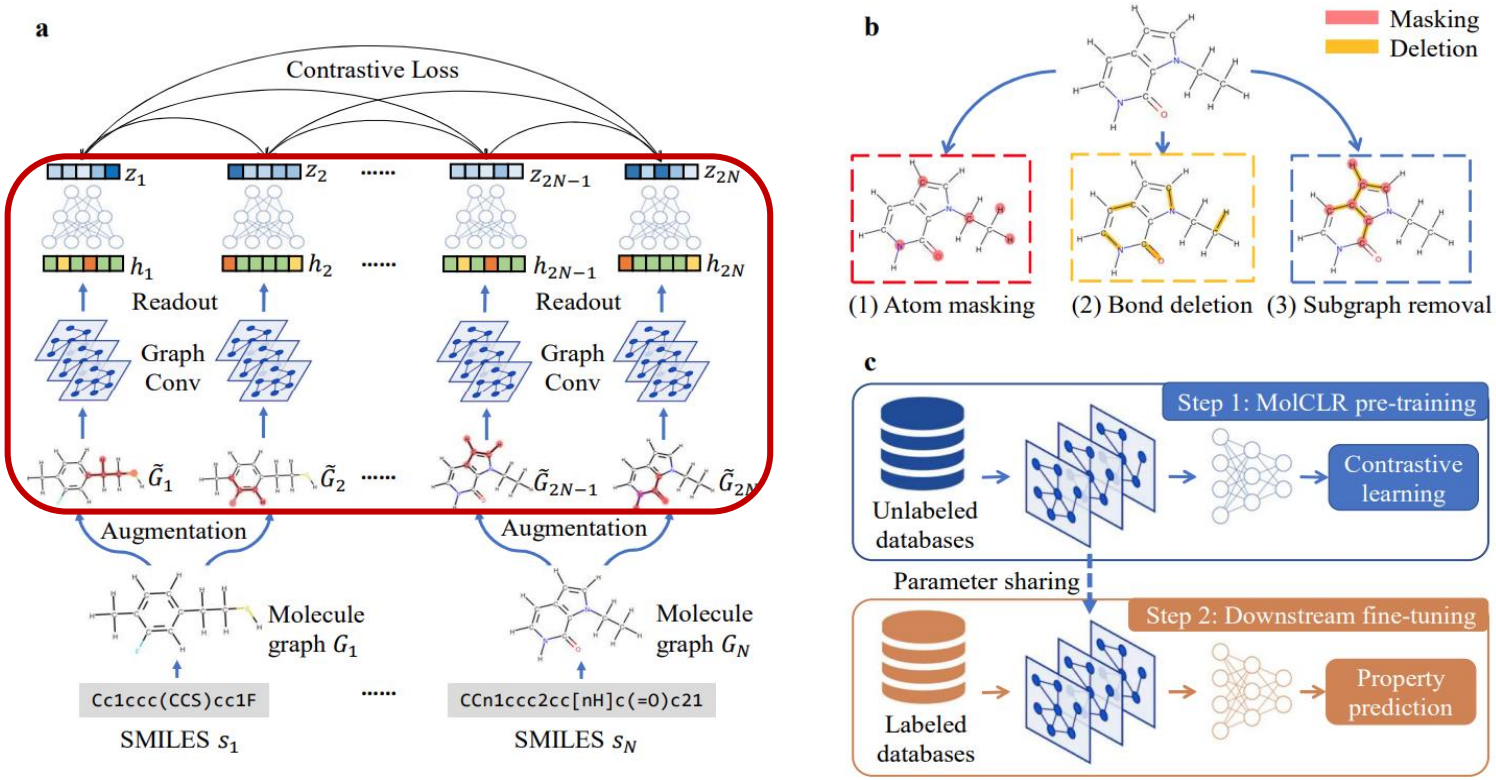
DMQA

MolCLR

Graph

❖ **Mol**ecular **C**ontrastive **L**earning of **R**epresentations (MolCLR)

- 분자 그래프 표현 학습을 위해 SimCLR 적용 – 분자 그래프 요약
  - ✓ Graph Convolutional Network (GCN), Graph Isomorphism Network (GIN) 인코더 적용
  - ✓ 그래프 인코더로 요약된 특징을 MLP로 한번 더 요약



DMQA

MolCLR

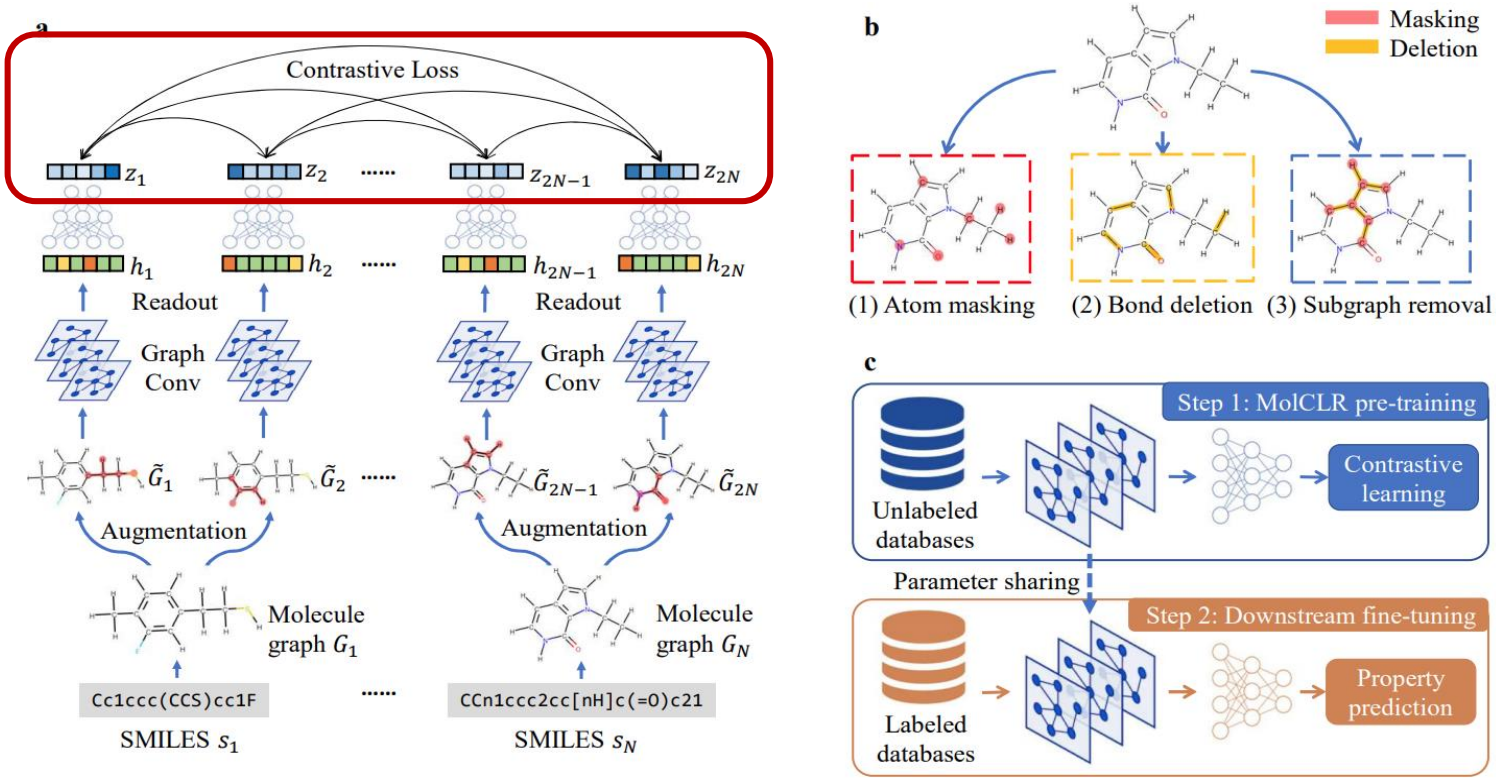
$$\mathcal{L}_i = -log\frac{exp\left(\frac{i \cdot i^+}{\tau}\right)}{exp\left(\frac{i \cdot i^+}{\tau}\right) + \sum_{k \notin \{i, i^+\}} exp\left(\frac{i \cdot k}{\tau}\right)}$$

i: Anchor
i^+: Positive
k: Negative Examples

Graph

❖ **Mol**ecular **C**ontrastive **L**earning of **R**epresentations (MolCLR)

- 분자 그래프 표현 학습을 위해 SimCLR 적용 – InfoNCE Loss Function
  - ✓ MLP로 요약된 벡터들을 사용하여 대조 학습 수행



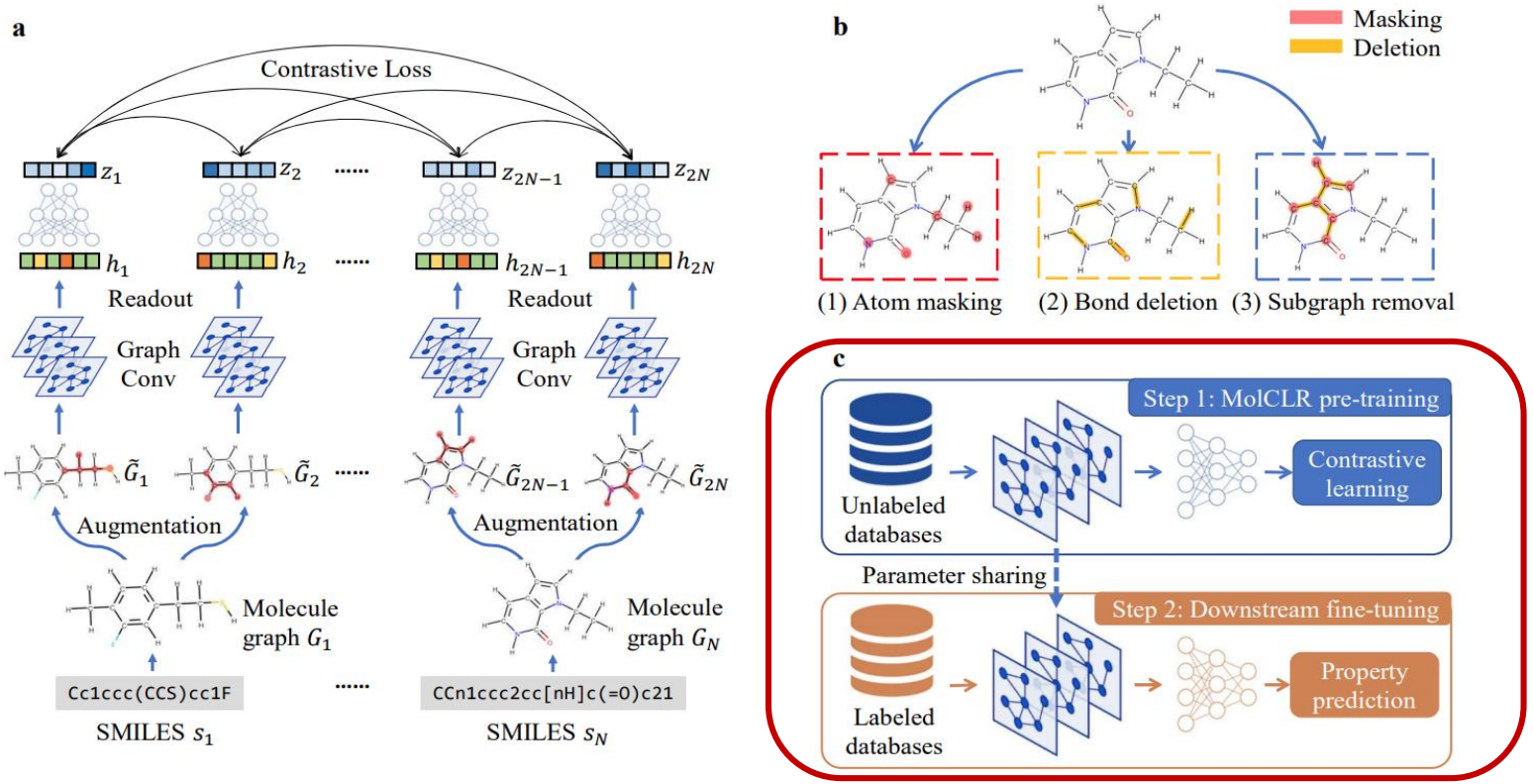
DMQA

MolCLR

Graph

❖ **Mol**ecular **C**ontrastive **L**earning of **R**epresentations (MolCLR)

- 분자 그래프 표현 학습을 위해 SimCLR 적용 – 전체 학습 프로세스
  - ✓ Step 1: 제안한 MolCLR 방법으로 분자 그래프 표현 학습을 수행(Pre-Training)
  - ✓ Step 2: 학습된 그래프 인코더를 사용하여 분자 성질 예측 수행(Downstream Tasks)



DMQA

Experiments

Graph

❖ Experiments – Molecular Property Predictions

- Seven Classification Benchmarks (MoleculeNet) 사용
- Supervised Learning / Self-Supervised or Pre-Training Method에 대한 결과
- ROC-AUC (%)로 평가

	Dataset	BBBP	Tox21	ClinTox	HIV	BACE	SIDER	MUV
	# Molecules	2039	7831	1478	41127	1513	1427	93087
	# Tasks	1	12	2	1	1	27	17
Supervised Learning	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
	GCN[17]	71.8±0.9	70.9±2.6	62.5±2.8	74.0±3.0	71.6±2.0	53.6±3.2	71.6±4.0
	GIN[18]	65.8±4.5	74.0±0.8	58.0±4.4	75.3±1.9	70.1±5.4	57.3±1.6	71.8±2.5
	SchNet[19]	84.8±2.2	77.2±2.3	71.5±3.7	70.2±3.4	76.6±1.1	53.9±3.7	71.3±3.0
	MGCN[52]	**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[20]	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**
Self-Supervised or Pre-Training	Hu et al.[45]	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[44]	**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
	$\text{MolCLR}_{\text{GCN}}$	73.8±0.2	74.7±0.8	86.7±1.0	77.8±0.5	78.8±0.5	66.9±1.2	84.0±1.8
	$\text{MolCLR}_{\text{GIN}}$	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**

Table 1. Test performance of different models on seven classification benchmarks. The first seven models are supervised learning methods and the last four are self-supervised/pre-training methods. Mean and standard deviation of test ROC-AUC (%) on each benchmark are reported.*

*Best performing supervised and self-supervised/pre-training methods for each benchmark are marked as bold.

Experiments

Graph

❖ Experiments – Molecular Property Predictions

- Six Regression Benchmarks (MoleculeNet) 사용
- Supervised Learning / Self-Supervised or Pre-Training Method에 대한 결과
- Root Mean Square Error (RMSE)로 평가

	Dataset	FreeSolv	ESOL	Lipo	QM7	QM8	QM9
	# Molecules	642	1128	4200	6830	21786	130829
	# Tasks	1	1	1	1	12	8
Supervised Learning	RF	**2.03±0.22**	1.07±0.19	0.88±0.04	122.7±4.2	0.0423±0.0021	16.061±0.019
	SVM	3.14±0.00	1.50±0.00	0.82±0.00	156.9±0.0	0.0543±0.0010	24.613±0.144
	GCN[17]	2.87±0.14	1.43±0.05	0.85±0.08	122.9±2.2	0.0366±0.0011	5.796±1.969
	GIN[18]	2.76±0.18	1.45±0.02	0.85±0.07	124.8±0.7	0.0371±0.0009	4.741±0.912
	SchNet[19]	3.22±0.76	1.05±0.06	0.91±0.10	**74.2±6.0**	0.0204±0.0021	0.081±0.001
	MGCN[52]	3.35±0.01	1.27±0.15	1.11±0.04	77.6±4.7	0.0223±0.0021	**0.050±0.002**
	D-MPNN[20]	2.18±0.91	**0.98±0.26**	**0.65±0.05**	105.8±13.2	**0.0143±0.0022**	3.241±0.119
Self-Supervised or Pre-Training	Hu et al.[45]	2.83±0.12	1.22±0.02	0.74±0.00	110.2±6.4	0.0191±0.0003	4.349±0.061
	N-Gram[44]	2.51±0.19	**1.10±0.03**	0.88±0.12	125.6±1.5	0.0320±0.0032	7.636±0.027
	$\text{MolCLR}_{\text{GCN}}$	2.39±0.14	1.16±0.00	0.78±0.01	**83.1±4.0**	0.0181±0.0002	3.552±0.041
	$\text{MolCLR}_{\text{GIN}}$	**2.20±0.20**	**1.11±0.01**	**0.65±0.08**	87.2±2.0	**0.0174±0.0013**	**2.357±0.118**

Table 2. Test performance of different models on six regression benchmarks. The first seven models are supervised learning methods and the last four are self-supervised/pre-training methods. Mean and standard deviation of test RMSE (for FreeSolv, ESOL, Lipo) or MAE (for QM7, QM8, QM9) are reported.*

*Best performing supervised and self-supervised/pre-training methods for each benchmark are marked as bold.

Conclusion

- ❖ 화학 분야 중 지도 학습 기반 분자 성질 예측, 신약 발견 분야에서 큰 성공을 거두었음
- ❖ 하지만, 레이블이 있는 데이터를 수집하는데 비용, 시간 소비가 매우 큼
- ❖ 레이블이 없는 데이터를 사용하여 분자 정보를 세밀하게 표현할 수 있는 대조 학습 제안
- ❖ 특히, 분자 구조를 그래프로 표현하는 그래프 기반 대조 학습 제안
- ❖ 여러 벤치마크에서 지도 학습 기반보다 제안하는 방법의 우수한 성능을 증명
  - 후기: 대조 학습을 전문적으로 연구하고 있는 사람으로서 다른 형태의 데이터에 Positive Pair, Negative Examples를 어떻게 정의하는지 알 수 있었고 이를 계기로 다양한 분야, 여러 데이터 형태에 대조 학습을 응용할 수 있을 것으로 생각됨

- ❖ Reference

  - Wang, Y., Wang, J., Cao, Z., & Farimani, A. B. (2021). MolCLR: molecular contrastive learning of representations via graph neural networks. arXiv preprint arXiv:2102.10056.