# Moleco: Molecolar Contrastive Learning with Chemical Language Models for Molecular Property Prediction

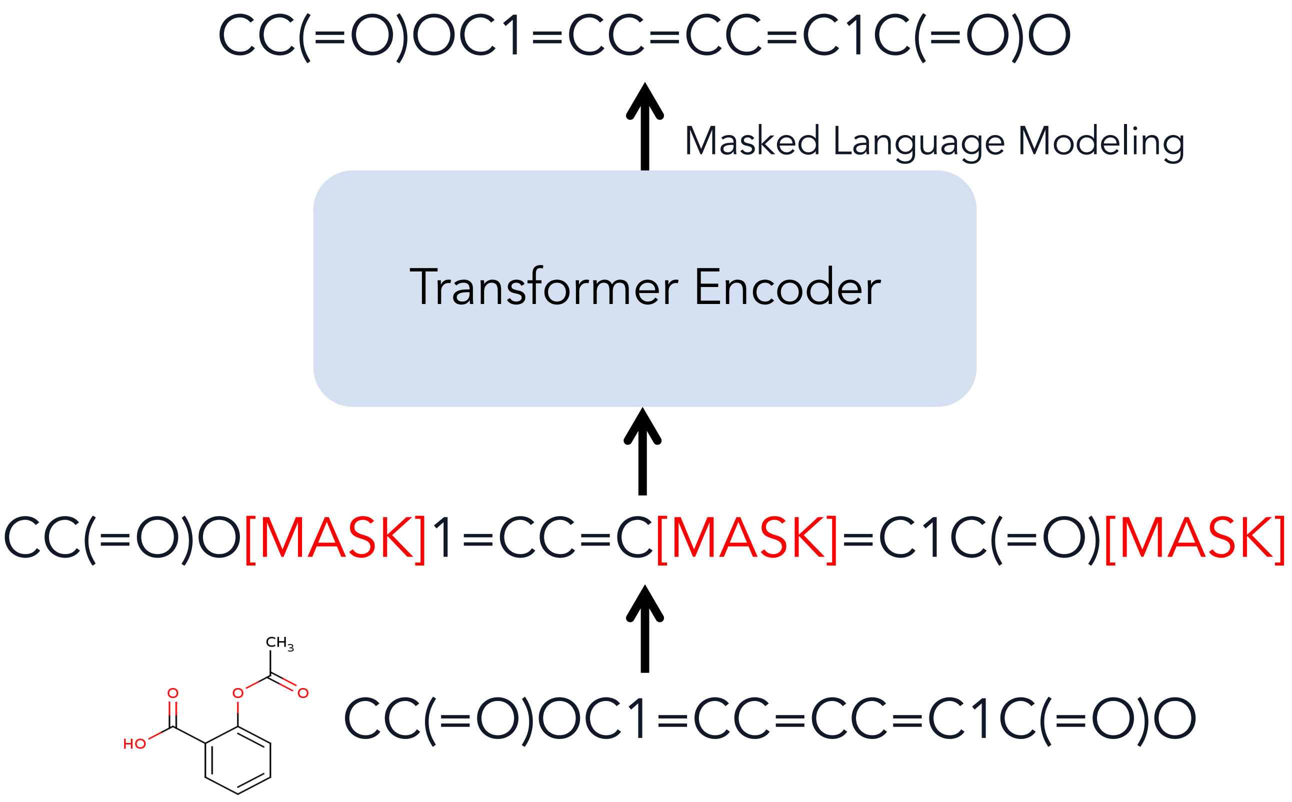
Jun-Hyung Park*, Hyuntae Park*, Yeachan Kim, Woosang Lim, SangKeun Lee

Data Intelligence Lab, Korea University
POSCO Holdings
pht0639@korea.ac.kr

2024.10.27

# Chemical Language Models (CLMs)

- CLMs are often trained with string-based descriptors, such as SMILES
- ChemBERTa, MoLFormer-XL, SELFormer, MolTRES, …



CC(=O)OC1=CC=CC=C1C(=O)O

Masked Language Modeling

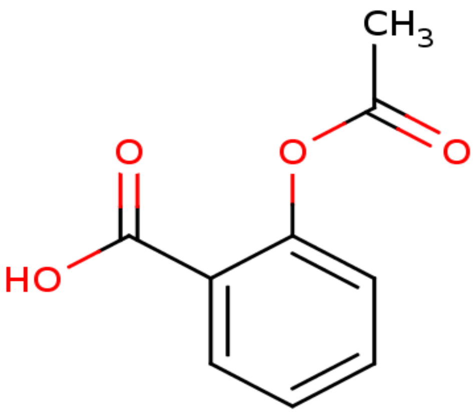
Transformer Encoder

CC(=O)O[MASK]1=CC=C[MASK]=C1C(=O)[MASK]

CC(=O)OC1=CC=CC=C1C(=O)O

# Chemical Language Models (CLMs)

- CLMs are often trained with string-based descriptors, such as SMILES
- However, SMILES implicitly contains limited structural information



Molecular structure

↔

CC(=O)OC1=CC=CC=C1C(=O)O

Linear representation

# Molecular Structure and Property

- Molecules with similar structures are likely to exhibit similar properties

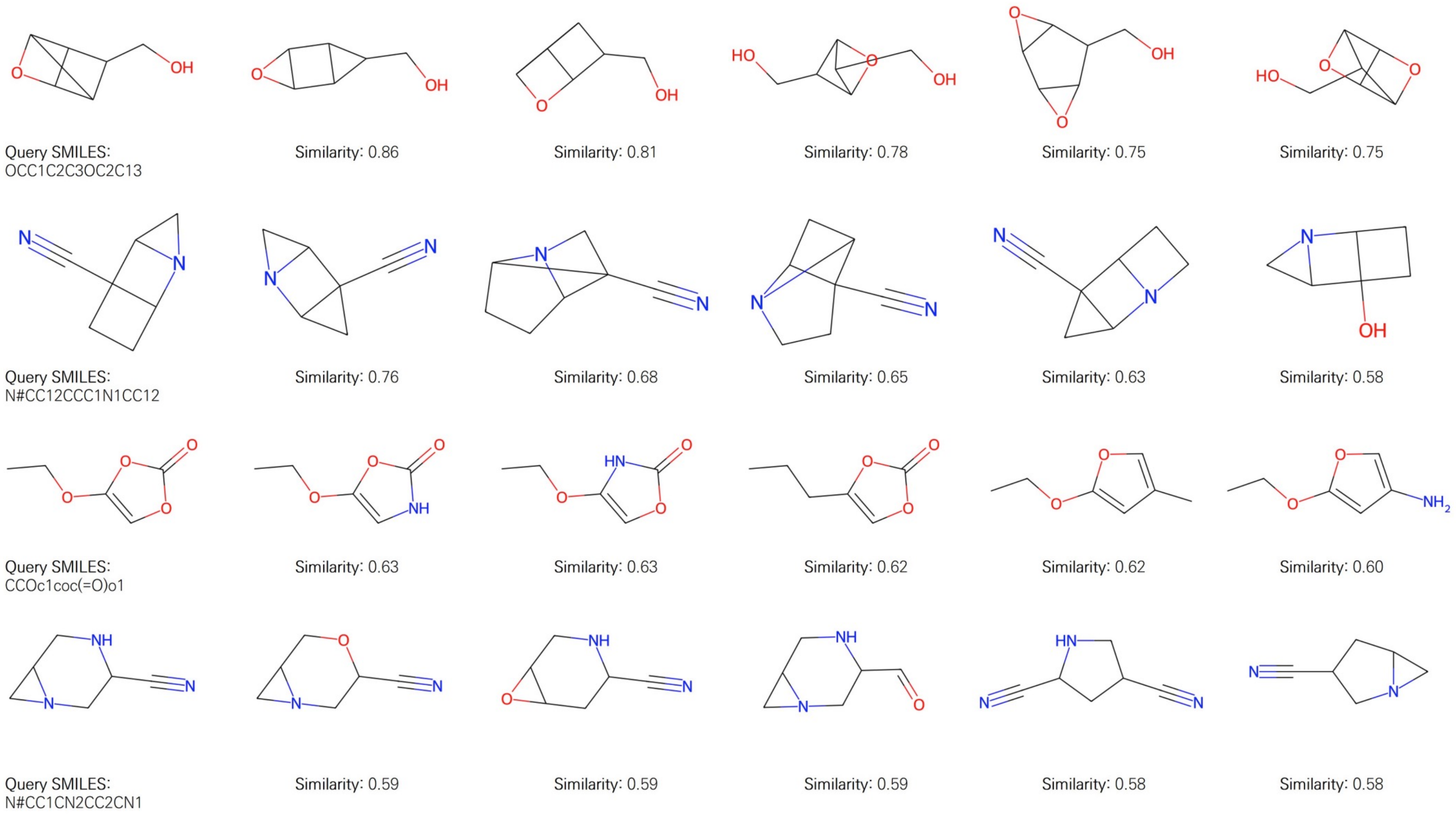


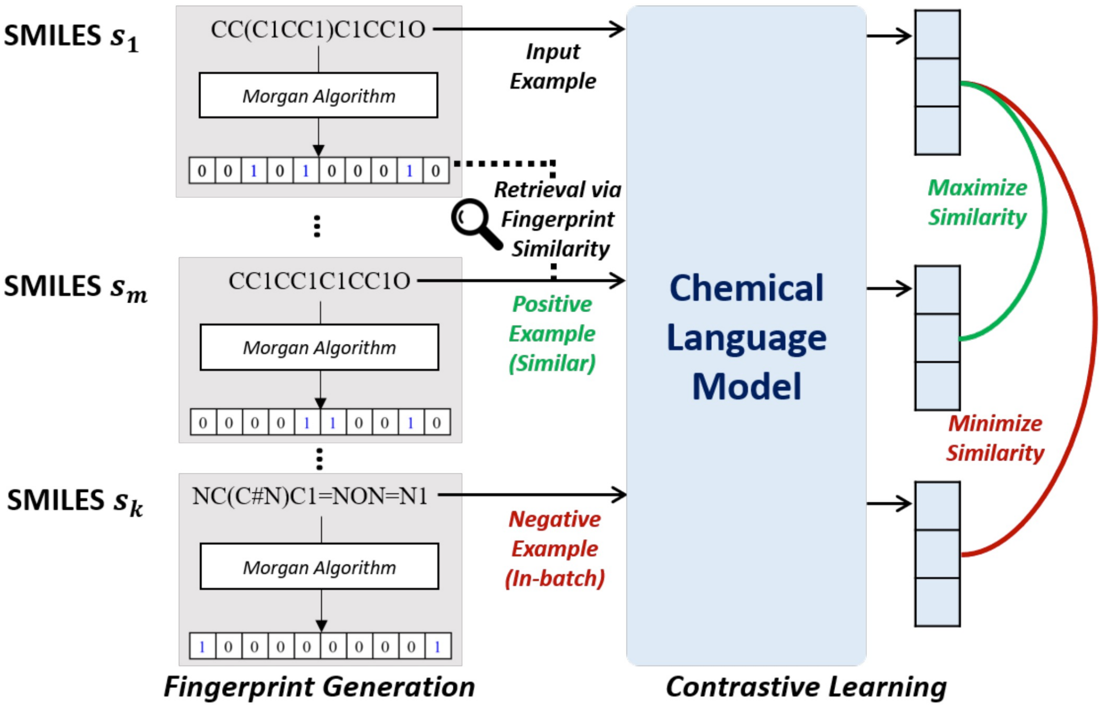
Figure 3: Visualization of the top pairs in the QM9 dataset.

# Research Question

How to enhance the understanding of CLMs
on the structural information?

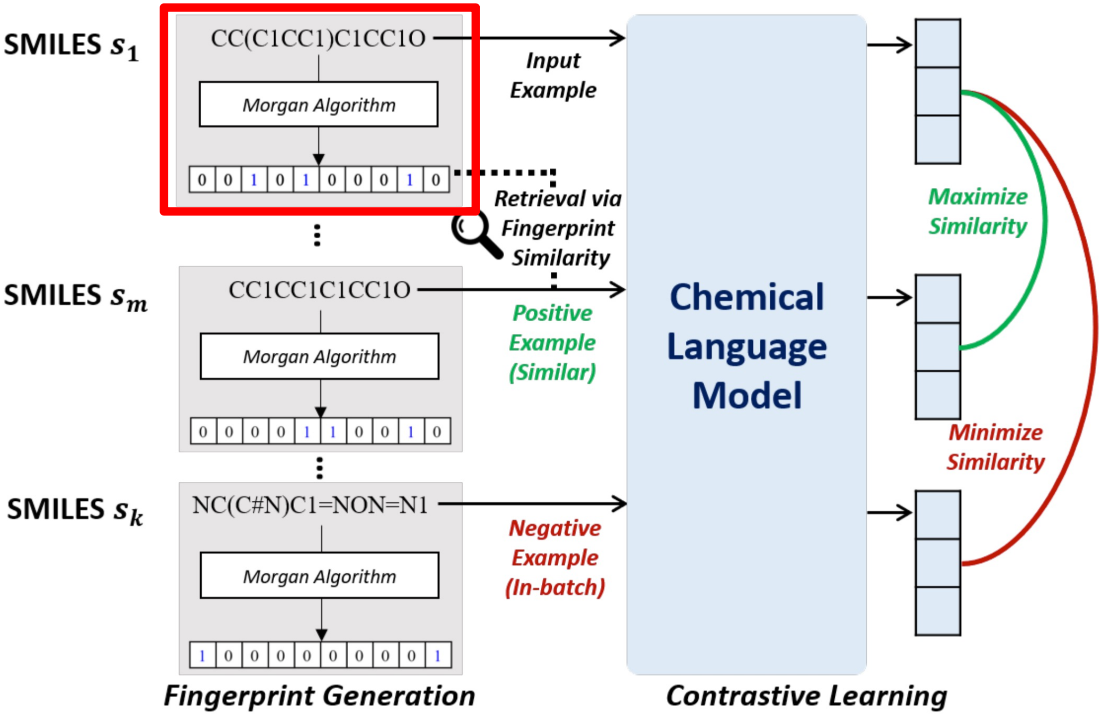
# Moleco

- Molecular Contrastive Learning with Chemical Language Models
- Combines fingerprint-based structural similarity with contrastive learning



# Moleco

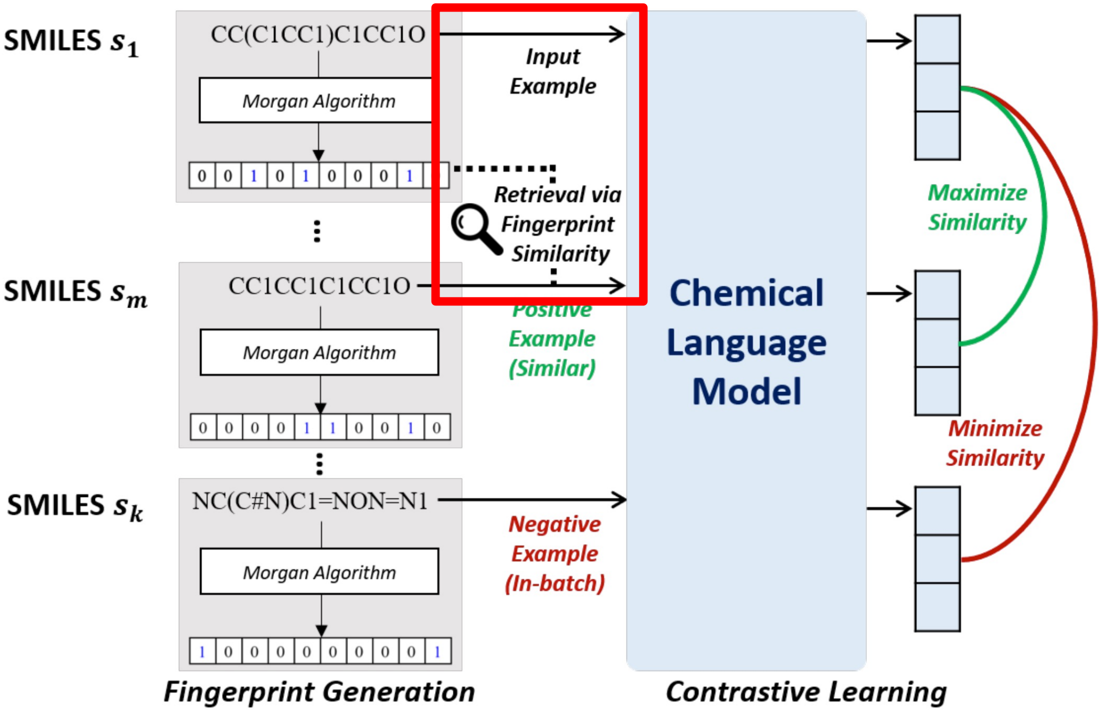
- Molecular Contrastive Learning with Chemical Language Models
- Combines fingerprint-based structural similarity with contrastive learning



- Extract fingerprints (ECFP4)

# Moleco

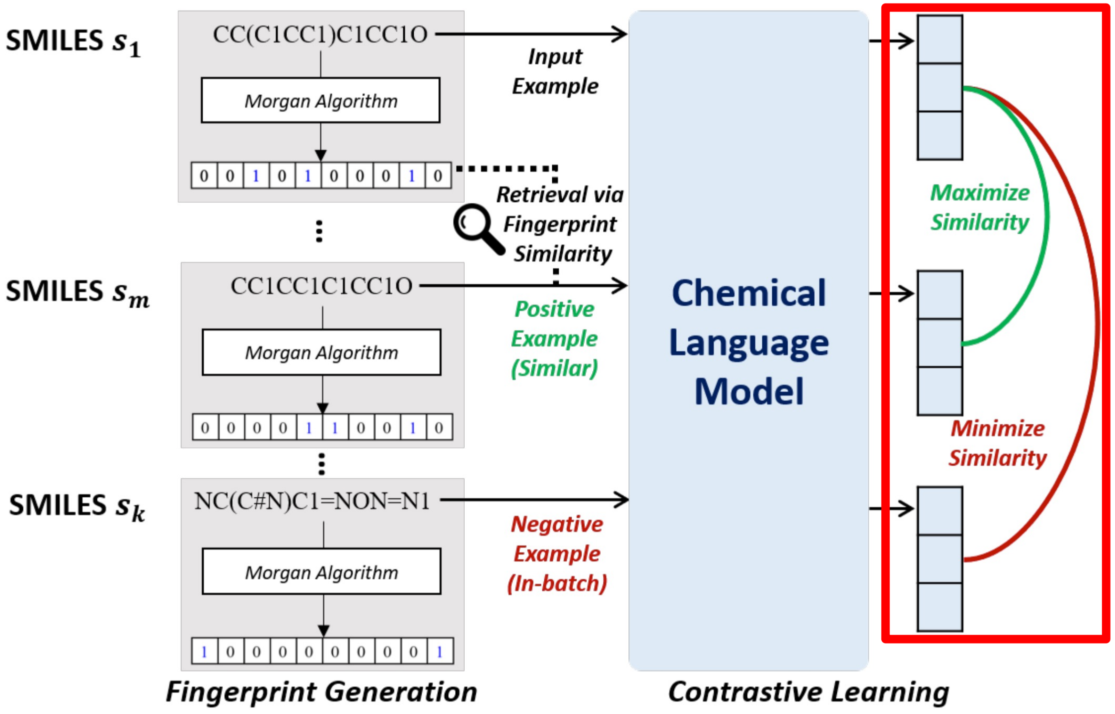
- Molecular Contrastive Learning with Chemical Language Models
- Combines fingerprint-based structural similarity with contrastive learning



- Extract fingerprints (ECFP4)
- Define similar molecules

# Moleco

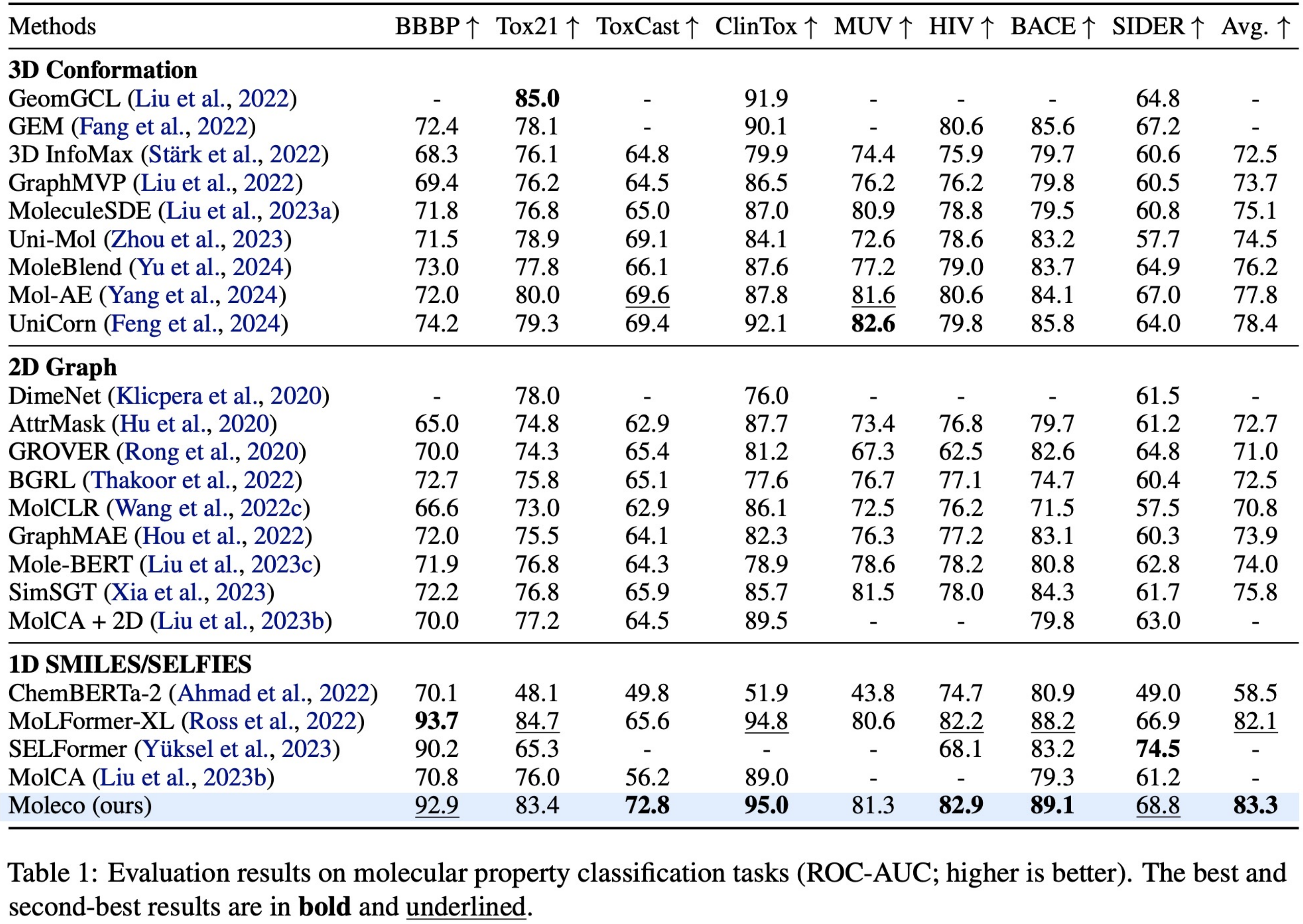
- Molecular Contrastive Learning with Chemical Language Models
- Combines fingerprint-based structural similarity with contrastive learning



- Extract fingerprints (ECFP4)
- Define similar molecules
- Distinguish between structurally similar and dissimilar molecules

# Experiments (MoleculeNet Classification)

- Contrasting structural similar molecules can improve performance

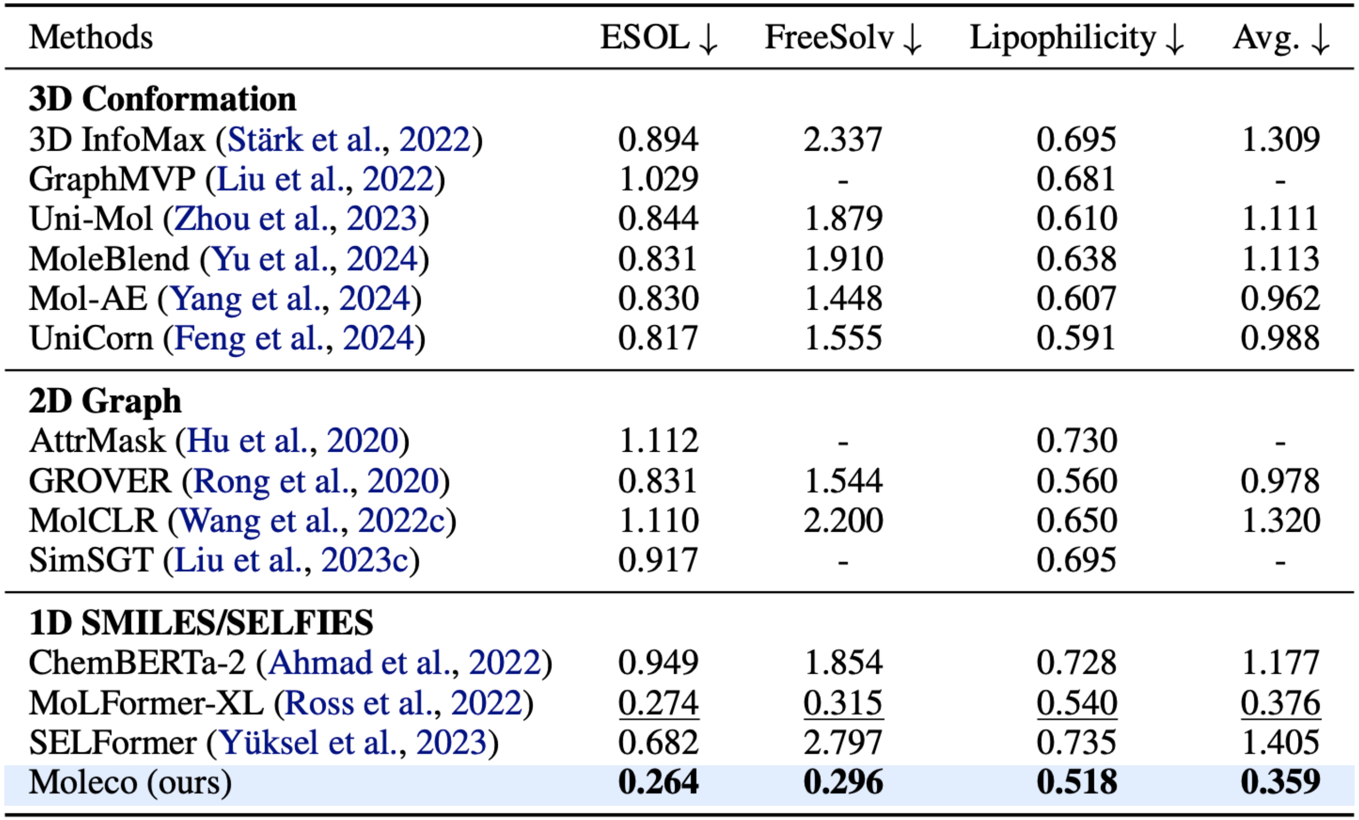


| Methods | BBBP ↑ | Tox21 ↑ | ToxCast ↑ | ClinTox ↑ | MUV ↑ | HIV ↑ | BACE ↑ | SIDER ↑ | Avg. ↑ |
|---|---|---|---|---|---|---|---|---|---|
| **3D Conformation** | | | | | | | | | |
| GeomGCL (Liu et al., 2022) | - | **85.0** | - | 91.9 | - | - | - | 64.8 | - |
| GEM (Fang et al., 2022) | 72.4 | 78.1 | - | 90.1 | - | 80.6 | 85.6 | 67.2 | - |
| 3D InfoMax (Stärk et al., 2022) | 68.3 | 76.1 | 64.8 | 79.9 | 74.4 | 75.9 | 79.7 | 60.6 | 72.5 |
| GraphMVP (Liu et al., 2022) | 69.4 | 76.2 | 64.5 | 86.5 | 76.2 | 76.2 | 79.8 | 60.5 | 73.7 |
| MoleculeSDE (Liu et al., 2023a) | 71.8 | 76.8 | 65.0 | 87.0 | 80.9 | 78.8 | 79.5 | 60.8 | 75.1 |
| Uni-Mol (Zhou et al., 2023) | 71.5 | 78.9 | 69.1 | 84.1 | 72.6 | 78.6 | 83.2 | 57.7 | 74.5 |
| MoleBlend (Yu et al., 2024) | 73.0 | 77.8 | 66.1 | 87.6 | 77.2 | 79.0 | 83.7 | 64.9 | 76.2 |
| Mol-AE (Yang et al., 2024) | 72.0 | 80.0 | 69.6 | 87.8 | 81.6 | 80.6 | 84.1 | 67.0 | 77.8 |
| UniCorn (Feng et al., 2024) | 74.2 | 79.3 | 69.4 | 92.1 | **82.6** | 79.8 | 85.8 | 64.0 | 78.4 |
| **2D Graph** | | | | | | | | | |
| DimeNet (Klicpera et al., 2020) | - | 78.0 | - | 76.0 | - | - | - | 61.5 | - |
| AttrMask (Hu et al., 2020) | 65.0 | 74.8 | 62.9 | 87.7 | 73.4 | 76.8 | 79.7 | 61.2 | 72.7 |
| GROVER (Rong et al., 2020) | 70.0 | 74.3 | 65.4 | 81.2 | 67.3 | 62.5 | 82.6 | 64.8 | 71.0 |
| BGRL (Thakoor et al., 2022) | 72.7 | 75.8 | 65.1 | 77.6 | 76.7 | 77.1 | 74.7 | 60.4 | 72.5 |
| MolCLR (Wang et al., 2022c) | 66.6 | 73.0 | 62.9 | 86.1 | 72.5 | 76.2 | 71.5 | 57.5 | 70.8 |
| GraphMAE (Hou et al., 2022) | 72.0 | 75.5 | 64.1 | 82.3 | 76.3 | 77.2 | 83.1 | 60.3 | 73.9 |
| Mole-BERT (Liu et al., 2023c) | 71.9 | 76.8 | 64.3 | 78.9 | 78.6 | 78.2 | 80.8 | 62.8 | 74.0 |
| SimSGT (Xia et al., 2023) | 72.2 | 76.8 | 65.9 | 85.7 | 81.5 | 78.0 | 84.3 | 61.7 | 75.8 |
| MolCA + 2D (Liu et al., 2023b) | 70.0 | 77.2 | 64.5 | 89.5 | - | - | 79.8 | 63.0 | - |
| **1D SMILES/SELFIES** | | | | | | | | | |
| ChemBERTa-2 (Ahmad et al., 2022) | 70.1 | 48.1 | 49.8 | 51.9 | 43.8 | 74.7 | 80.9 | 49.0 | 58.5 |
| MoLFormer-XL (Ross et al., 2022) | **93.7** | 84.7 | 65.6 | 94.8 | 80.6 | 82.2 | 88.2 | 66.9 | 82.1 |
| SELFormer (Yüksel et al., 2023) | 90.2 | 65.3 | - | - | - | 68.1 | 83.2 | **74.5** | - |
| MolCA (Liu et al., 2023b) | 70.8 | 76.0 | 56.2 | 89.0 | - | - | 79.3 | 61.2 | - |
| Moleco (ours) | 92.9 | 83.4 | **72.8** | **95.0** | 81.3 | **82.9** | **89.1** | 68.8 | **83.3** |

Table 1: Evaluation results on molecular property classification tasks (ROC-AUC; higher is better). The best and second-best results are in **bold** and underlined.

# Experiments (MoleculeNet Regression)

- Contrasting structural similar molecules can improve performance

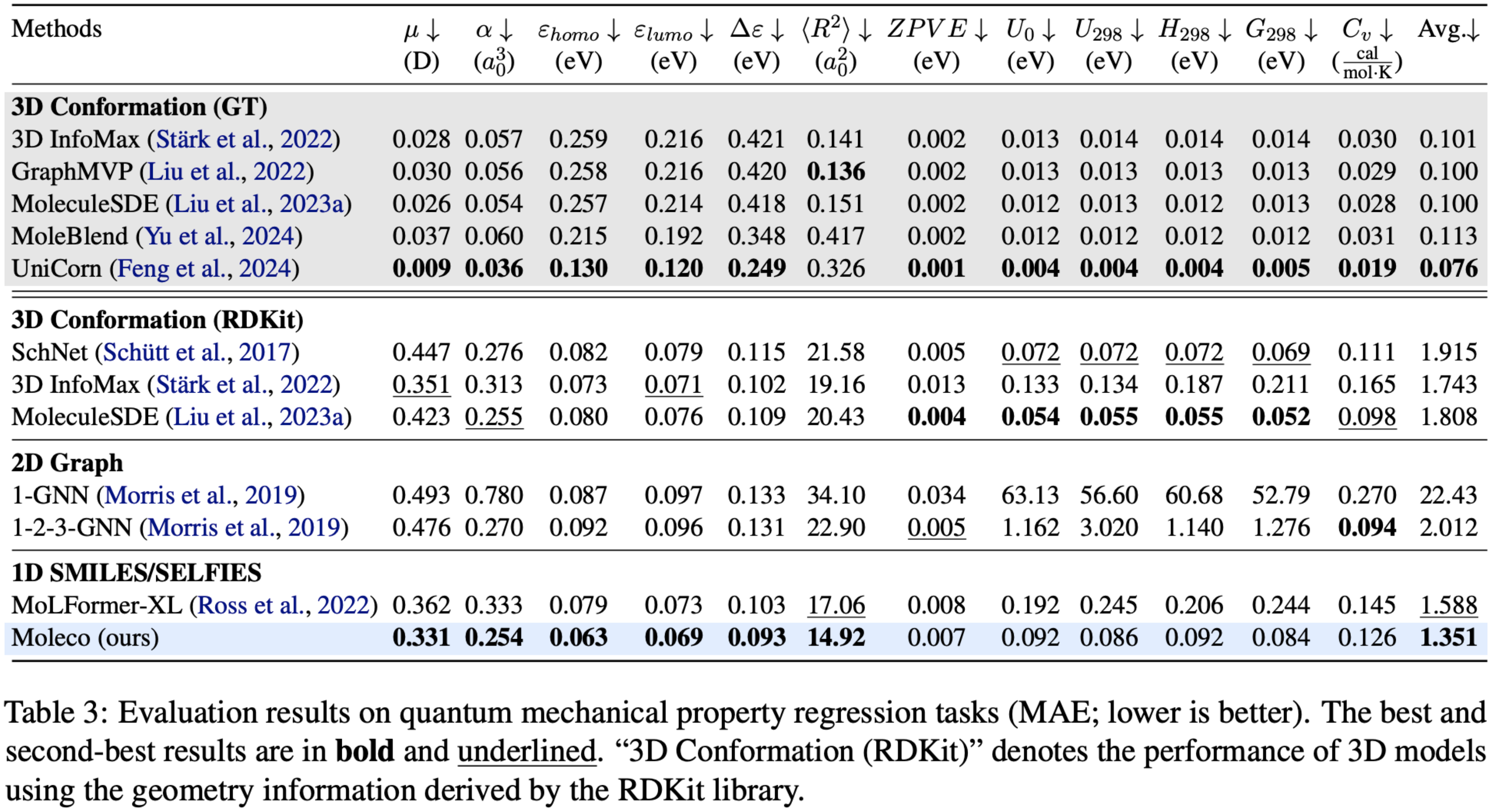


| Methods | ESOL ↓ | FreeSolv ↓ | Lipophilicity ↓ | Avg. ↓ |
|---|---|---|---|---|
| **3D Conformation** | | | | |
| 3D InfoMax (Stärk et al., 2022) | 0.894 | 2.337 | 0.695 | 1.309 |
| GraphMVP (Liu et al., 2022) | 1.029 | - | 0.681 | - |
| Uni-Mol (Zhou et al., 2023) | 0.844 | 1.879 | 0.610 | 1.111 |
| MoleBlend (Yu et al., 2024) | 0.831 | 1.910 | 0.638 | 1.113 |
| Mol-AE (Yang et al., 2024) | 0.830 | 1.448 | 0.607 | 0.962 |
| UniCorn (Feng et al., 2024) | 0.817 | 1.555 | 0.591 | 0.988 |
| **2D Graph** | | | | |
| AttrMask (Hu et al., 2020) | 1.112 | - | 0.730 | - |
| GROVER (Rong et al., 2020) | 0.831 | 1.544 | 0.560 | 0.978 |
| MolCLR (Wang et al., 2022c) | 1.110 | 2.200 | 0.650 | 1.320 |
| SimSGT (Liu et al., 2023c) | 0.917 | - | 0.695 | - |
| **1D SMILES/SELFIES** | | | | |
| ChemBERTa-2 (Ahmad et al., 2022) | 0.949 | 1.854 | 0.728 | 1.177 |
| MoLFormer-XL (Ross et al., 2022) | 0.274 | 0.315 | 0.540 | 0.376 |
| SELFormer (Yüksel et al., 2023) | 0.682 | 2.797 | 0.735 | 1.405 |
| Moleco (ours) | **0.264** | **0.296** | **0.518** | **0.359** |

Table 2: Evaluation results on molecular property regression tasks (RMSE; lower is better). The best and second-best results are in **bold** and underlined.

# Experiments (QM9)

- Moleco can provide accurate prediction of quantum properties without ground-truth geometry information



| Methods | $\mu \downarrow$ (D) | $\alpha \downarrow$ ($a_0^3$) | $\varepsilon_{homo} \downarrow$ (eV) | $\varepsilon_{lumo} \downarrow$ (eV) | $\Delta\varepsilon \downarrow$ (eV) | $\langle R^2 \rangle \downarrow$ ($a_0^2$) | $ZPVE \downarrow$ (eV) | $U_0 \downarrow$ (eV) | $U_{298} \downarrow$ (eV) | $H_{298} \downarrow$ (eV) | $G_{298} \downarrow$ (eV) | $C_v \downarrow$ ($\frac{\text{cal}}{\text{mol}\cdot\text{K}}$) | Avg.↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **3D Conformation (GT)** | | | | | | | | | | | | | |
| 3D InfoMax (Stärk et al., 2022) | 0.028 | 0.057 | 0.259 | 0.216 | 0.421 | 0.141 | 0.002 | 0.013 | 0.014 | 0.014 | 0.014 | 0.030 | 0.101 |
| GraphMVP (Liu et al., 2022) | 0.030 | 0.056 | 0.258 | 0.216 | 0.420 | **0.136** | 0.002 | 0.013 | 0.013 | 0.013 | 0.013 | 0.029 | 0.100 |
| MoleculeSDE (Liu et al., 2023a) | 0.026 | 0.054 | 0.257 | 0.214 | 0.418 | 0.151 | 0.002 | 0.012 | 0.013 | 0.012 | 0.013 | 0.028 | 0.100 |
| MoleBlend (Yu et al., 2024) | 0.037 | 0.060 | 0.215 | 0.192 | 0.348 | 0.417 | 0.002 | 0.012 | 0.012 | 0.012 | 0.012 | 0.031 | 0.113 |
| UniCorn (Feng et al., 2024) | **0.009** | **0.036** | **0.130** | **0.120** | **0.249** | 0.326 | **0.001** | **0.004** | **0.004** | **0.004** | **0.005** | **0.019** | **0.076** |
| **3D Conformation (RDKit)** | | | | | | | | | | | | | |
| SchNet (Schütt et al., 2017) | 0.447 | 0.276 | 0.082 | 0.079 | 0.115 | 21.58 | 0.005 | 0.072 | 0.072 | 0.072 | 0.069 | 0.111 | 1.915 |
| 3D InfoMax (Stärk et al., 2022) | 0.351 | 0.313 | 0.073 | 0.071 | 0.102 | 19.16 | 0.013 | 0.133 | 0.134 | 0.187 | 0.211 | 0.165 | 1.743 |
| MoleculeSDE (Liu et al., 2023a) | 0.423 | 0.255 | 0.080 | 0.076 | 0.109 | 20.43 | **0.004** | **0.054** | **0.055** | **0.055** | **0.052** | 0.098 | 1.808 |
| **2D Graph** | | | | | | | | | | | | | |
| 1-GNN (Morris et al., 2019) | 0.493 | 0.780 | 0.087 | 0.097 | 0.133 | 34.10 | 0.034 | 63.13 | 56.60 | 60.68 | 52.79 | 0.270 | 22.43 |
| 1-2-3-GNN (Morris et al., 2019) | 0.476 | 0.270 | 0.092 | 0.096 | 0.131 | 22.90 | 0.005 | 1.162 | 3.020 | 1.140 | 1.276 | **0.094** | 2.012 |
| **1D SMILES/SELFIES** | | | | | | | | | | | | | |
| MoLFormer-XL (Ross et al., 2022) | 0.362 | 0.333 | 0.079 | 0.073 | 0.103 | 17.06 | 0.008 | 0.192 | 0.245 | 0.206 | 0.244 | 0.145 | 1.588 |
| Moleco (ours) | **0.331** | **0.254** | **0.063** | **0.069** | **0.093** | **14.92** | 0.007 | 0.092 | 0.086 | 0.092 | 0.084 | 0.126 | **1.351** |

Table 3: Evaluation results on quantum mechanical property regression tasks (MAE; lower is better). The best and second-best results are in **bold** and underlined. “3D Conformation (RDKit)” denotes the performance of 3D models using the geometry information derived by the RDKit library.

# Conclusion

- We propose Moleco, a novel contrastive learning framework that enhances CLM’s understanding of molecular structures

# Conclusion

- We propose Moleco, a novel contrastive learning framework that enhances CLM's understanding of molecular structures

- We develop a novel scheme to identify and leverage structurally similar molecules based on fingerprint-based structural similarity

# Conclusion

- We propose Moleco, a novel contrastive learning framework that enhances CLM’s understanding of molecular structures

- We develop a novel scheme to identify and leverage structurally similar molecules based on fingerprint-based structural similarity

- We verify that Moleco establishes new state-of-the-art results across a wide range of molecular property prediction tasks.

# Thanks !!!