EMNLP 2024

Molecular Contrastive Learning with Chemical Language Models for Molecular Property Prediction

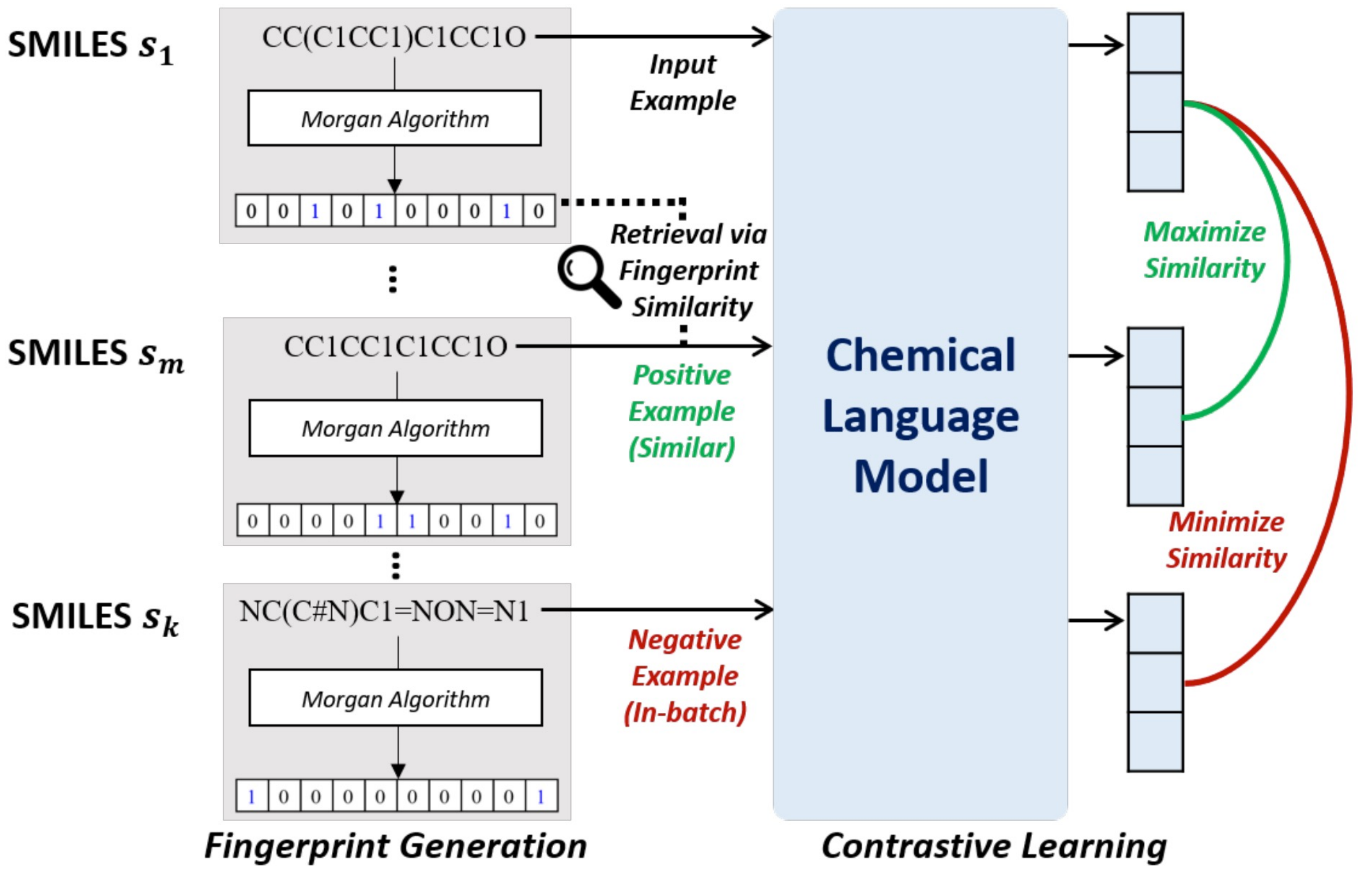
Jun-Hyung Park*, Hyuntae Park*, Yeachan Kim, Woosang Lim, SangKeun Lee

Korea University, POSCO Holdings

Introduction

- String-based descriptors such as SMILES capture molecular structures implicitly, limiting their utility for molecular property prediction where explicit structural information is essential.
- Current pre-training methods for chemical language models (CLMs) lack sufficient structural guidance, reducing their accuracy in associating structure with properties
- Moleco framework, based on fingerprint-derived structural similarities, enhances CLMs' ability to leverage structural details for better property prediction.

Our Framework (Moleco)



Experimental Results (Classification)

Methods	ToxCast ↑	ClinTox ↑	HIV ↑	BACE ↑	SIDER ↑	Avg. ↑
3D Conformation						
3D InfoMax (Stärk et al., 2022)	64.8	79.9	75.9	79.7	60.6	72.5
GraphMVP (Liu et al., 2022)	64.5	86.5	76.2	79.8	60.5	73.7
Uni-Mol (Zhou et al., 2023)	69.1	84.1	78.6	83.2	57.7	74.5
MoleBlend (Yu et al., 2024)	66.1	87.6	79.0	83.7	64.9	76.2
Mol-AE (Yang et al., 2024)	69.6	87.8	80.6	84.1	67.0	77.8
UniCorn (Feng et al., 2024)	69.4	92.1	79.8	85.8	64.0	78.4
2D Graph						
AttrMask (Hu et al., 2020)	62.9	87.7	76.8	79.7	61.2	72.7
GROVER (Rong et al., 2020)	65.4	81.2	62.5	82.6	64.8	71.0
MolCLR (Wang et al., 2022c)	62.9	86.1	76.2	71.5	57.5	70.8
SimSGT (Xia et al., 2023)	65.9	85.7	78.0	84.3	61.7	75.8
1D SMILES/SELFIES						
ChemBERTa-2 (Ahmad et al., 2022)	49.8	51.9	74.7	80.9	49.0	58.5
MoLFormer-XL (Ross et al., 2022)	65.6	94.8	82.2	88.2	66.9	82.1
SELFormer (Yüksel et al., 2023)	-	-	68.1	83.2	**74.5**	-
Moleco (ours)	**72.8**	**95.0**	**82.9**	**89.1**	68.8	**83.3**

Experimental Results (Regression)

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**

Experimental Results (QM9)

Methods	$\mu\downarrow$	$\alpha\downarrow$	$\varepsilon_{homo}\downarrow$	$\varepsilon_{lumo}\downarrow$	$\Delta\varepsilon\downarrow$	$\langle R^2\rangle\downarrow$	$ZPVE\downarrow$	$U_0\downarrow$	$U_{298}\downarrow$	$H_{298}\downarrow$	$G_{298}\downarrow$	$C_v\downarrow$	Avg.↓
	(D)	(a_0^3)	(eV)	(eV)	(eV)	(a_0^2)	(eV)	(eV)	(eV)	(eV)	(eV)	$(\frac{\text{cal}}{\text{mol}\cdot\text{K}})$	
3D Conformation													
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													
1-2-3-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**

Comparison of Moleco Variants

Backbone	Embeddings	Similarity	CLS ↑	REG ↓
MoLFormer-XL	Morgan FP	Cosine	**83.3**	**0.359**
	Morgan FP	Tanimoto	82.3	0.374
	Torsion FP	Cosine	82.0	0.383
	RDKit FP	Cosine	81.6	0.380
	3D GeoFormer	Cosine	80.6	0.379
ChemBERTa-2	MorganFP	Cosine	60.2	1.107

Conclusion & Takeaway

- **Impact of Moleco Framework**: Moleco enhances CLMs' understanding of molecular structures by employing contrastive learning, leading to improved molecular property prediction.
- **Effectiveness of Fingerprint-Based Similarity**: Moleco leverages fingerprint-based molecular similarities to identify relevant molecular pairs, showing significant improvements over state-of-the-art methods.
- **Performance Gains in Property Prediction**: Moleco achieves consistent performance gains across diverse molecular property prediction tasks, underscoring the importance of explicitly incorporating molecular structural information.
- **Takeaway (Importance of Molecular Structural Similarity)**: Incorporating molecular structural similarity through contrastive learning is crucial for enhancing CLMs' accuracy in molecular property prediction.