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

Jun-Hyung Park[1]*, **Hyuntae Park**[2*], **Yeachan Kim**[2], **Woosang Lim**[3], **SangKeun Lee**[2,4]
[1]Division of Language & AI, Hankuk University of Foreign Studies
[2]Department of Artificial Intelligence, Korea University
[3]POSCO Holdings
[4]Department of Computer Science and Engineering, Korea University
jhp@hufs.ac.kr {pht0639, yeachan, yalphy}@korea.ac.kr woosang_lim@posco-inc.com

Abstract

Pre-trained chemical language models (CLMs) excel in the field of molecular property prediction, utilizing string-based molecular descriptors such as SMILES for learning universal representations. However, such string-based descriptors implicitly contain limited structural information, which is closely associated with molecular property prediction. In this work, we introduce Moleco, a novel contrastive learning framework to enhance the understanding of molecular structures within CLMs. Based on the similarity of fingerprint vectors among different molecules, we train CLMs to distinguish structurally similar and dissimilar molecules in a contrastive manner. Experimental results demonstrate that Moleco significantly improves the molecular property prediction performance of CLMs, outperforming state-of-the-art models. Moreover, our in-depth analysis with diverse Moleco variants verifies that fingerprint vectors are highly effective features in improving CLMs' understanding of the structural information of molecules[1].

1 Introduction

In drug discovery and materials science, applying deep neural networks to molecular property prediction has brought increasing attention (Butler et al., 2018). These networks can predict molecular properties with a significantly reduced cost compared with traditional methods like wet lab experiments. Moreover, combined with transfer learning approaches with large-scale pre-training, deep neural networks have shown their versatility and generalization capacity, allowing for applying a single model across various tasks. This also reduces the need for task-specific modeling, leading to high usability and practicality.

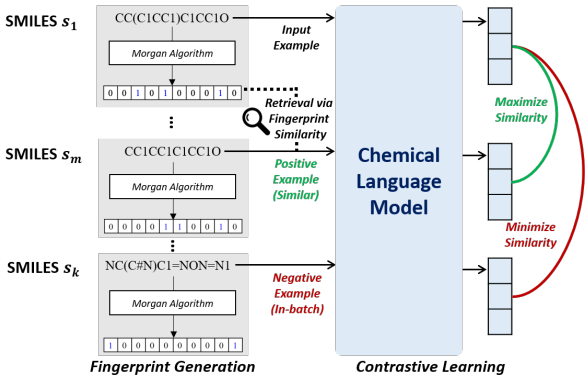

Figure 1: Illustration of Moleco. We extract and construct a set of similar molecules by measuring the cosine similarity between fingerprint vectors of different molecules. We subsequently maximize the agreement between pairs of structurally similar molecules, while minimizing that of the other molecules in a batch.

Recently, inspired by the success of the pre-trained language models (Devlin et al., 2019; Liu et al., 2019), chemical language models (CLMs) have been introduced and shown their excellence in predicting molecular properties (Chithrananda et al., 2020; Ahmad et al., 2022; Ross et al., 2022). These CLMs, typically employing Transformer architectures (Vaswani et al., 2017), are trained on large-scale string-based molecular descriptors to learn universal molecular representations. String-based molecular descriptors, such as Simplified Molecular-Input Line-Entry System (SMILES) (Weininger, 1988), compactly represent molecules in a text format, providing benefits in handling large-scale molecule data. Moreover, the employed Transformer architectures have shown high efficiency and parallelizability in processing large-scale molecular data.

Despite the efficiency of string-based molecular descriptors, they contain limited structural information of molecules in an implicit manner, which is critical in predicting molecular properties (Soares

* These authors contributed equally to this work.
[1]Our code and data are available at https://github.com/Park-ing-lot/Moleco

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 408–420
November 12-16, 2024

et al., 2023). For example, SMILES involves redundant, invalid descriptors of molecular structures and needs to be interpreted to uncover its structural information. Moreover, typical pre-training approaches for CLMs, namely masked language modeling (Devlin et al., 2019), do not explicitly train models to capture such structural information. Thus, current CLMs suffer from capturing the relationships between molecular structures and properties (Graff et al., 2023).

In this work, we introduce Moleco (***Mole**cular **Co**ntrastive Learning with Chemical Language Models*), a novel contrastive learning framework to enhance the understanding of CLMs on the structural information of molecules. Moleco leverages contrastive learning among different molecules based on a structural similarity calculated using fingerprint embeddings (Rogers and Hahn, 2010), which contain substructure information of molecules. Specifically, based on the fingerprint embeddings, Moleco identifies and utilizes top-k structurally similar molecules as positive samples, while using the others in batch as negative samples. This contrastive learning approach enriches models' representation to better reflect the relationships between different molecules in molecular substructures. Furthermore, Moleco additionally trains CLMs with a prediction of structural embeddings in a multitask learning manner, as direct guidance of structure information of molecules.

We evaluate Moleco on various tasks from MoleculeNet benchmarks (Wu et al., 2018), including eight classification and four regression tasks. Our extensive experiments verify that, although fingerprints are highly simplified and straightforward methods to embed structural information of molecules, Moleco significantly improves the molecular property prediction of CLMs. Notably, Moleco achieves performance improvements of 1.8% and 7.3% on average in molecular property classification and regression tasks, respectively, compared with state-of-the-art models. Moreover, our in-depth analysis demonstrates that the proposed fingerprint-based similarity effectively identifies structurally similar molecules, leading to the improvements in CLMs' understanding of structural properties of molecules.

Our main contributions are as follows:

- We propose Moleco, a novel contrastive learning framework that enhances CLMs' 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.

2 Related Work

2.1 Chemical Language Models

Self-supervised learning, with its substantial success in various research domains, has inspired numerous works on molecular property prediction. Recently, inspired by the development of Natural Language Processing (Devlin et al., 2019), string-based molecular descriptors such as SMILES (Weininger, 1988) and SELFIES (Krenn et al., 2022) have been utilized to learn universal molecular representations with Transformer architecture (Wang et al., 2023a; Ross et al., 2022; Yüksel et al., 2023). Particularly, Ross et al. (2022) have achieved superior performance on molecular property predictions by learning universal molecular representations with 1.2 billion SMILES sequences. Yüksel et al. (2023) have proposed SELFormer, a string-based Transformer architecture model that utilizes SELFIES, aimed at learning robust molecular representations. Due to the extensive quantity of data, these approaches have achieved significant performance improvements in molecular property prediction. However, these methods do not involve an explicit scheme to capture the complete structural information of molecules.

2.2 Chemical Graph Models

Another line of work (You et al., 2020; Wang et al., 2022a,b; Rong et al., 2020; Zang et al., 2023) has focused on learning molecular representations with 2D topology information of molecules, since a graph is a natural representation of molecules and conveys structural information. Rong et al. (2020); Zang et al. (2023) have proposed to pre-train GNN or Transformer models with a self-supervised learning method on graphs to learn rich structural and semantic information of molecules. In addition, pre-training models with 3D geometry information have been proposed to boost molecular property prediction (Stärk et al., 2022; Fang et al., 2022; Liu et al., 2022). Fang et al. (2022) have proposed GEM, a self-supervised framework using molecular geometric information. Liu et al. (2022)

has conducted fragment-based contrastive learning with geometric inputs. Distantly related to our framework, these methods utilize 2D or 3D graphs including explicit structural information to represent molecules, coming with a higher complexity.

2.3 Fingerprint-based Chemical Models

Meanwhile, several works have leveraged molecular fingerprints in diverse molecular tasks. Fingerprints such as ECFPs (Rogers and Hahn, 2010) have been developed to encode structural information of molecules into binary vectors for similarity searching. Earlier machine learning approaches (Cereto-Massagué et al., 2015; Coley et al., 2017) learned molecular representation from fingerprints. Kuang et al. (2024) have pre-trained a model with contrastive learning based on 3D conformation descriptors and ECFPs to figure out positive and negative examples. Zhu et al. (2022) have proposed MEMO that utilizes different molecular featurization techniques, including 2D topology, 3D geometry, SMILES string, and fingerprint, to obtain a better representation of molecules. In this work, we leverage the fingerprints to alleviate the limitations of string-based Transformer.

3 Methodology

In this section, we propose Moleco, which trains CLMs to explicitly learn structural similarities of different molecules in a contrastive manner. Specifically, we first obtain structural similarities using fingerprint embeddings of molecules, and subsequently train models to contrastively learn the similarities, as illustrated in Figure 1. In addition, we introduce an auxiliary training objective that directly predicts molecular embeddings, to further enhance the structural understanding of CLMs. After Moleco training, we fine-tune the models on downstream tasks to predict molecular properties.

3.1 Molecular Contrastive Learning

Understanding molecular structure-property relationships is crucial for accurately predicting functional outcomes, such as reactivity, stability, and biological activity (Le et al., 2012), since the molecules with similar structures often exhibit similar properties (Martin et al., 2002). To supplement CLMs' understanding of such relationships, we introduce Moleco, a novel molecular contrastive learning framework for CLMs. We train models to distinguish between structurally similar and dissimilar molecules in a contrastive manner. This approach is expected to facilitate the model's ability to determine properties by recognizing structural differences in molecules.

To this end, we employ fingerprints (Rogers and Hahn, 2010), multi-dimensional binary vectors describing the existence of particular substructures in a molecule, which can address the limitations of string-based descriptors utilized by CLMs. Specifically, we first create a set of structurally similar molecules for each molecule, denoted as H, identified by a similarity metric based on fingerprints. We extract 2048-dimensional fingerprints from the SMILES descriptor of each molecule based on the Morgan algorithm using the RDKit library[2]. By calculating the cosine similarity between these vectors, we identify the top-k similar molecules for each molecule. Subsequently, we sample a batch of N molecules and define the contrastive prediction task on pairs of similar molecules. For each molecule in a batch, we randomly select a molecule from the pre-identified set of similar molecules H to form the positive pair, resulting in $2N$ molecules in a final batch.

We then define the agreement between two molecule m_i and m_j in a batch as follows:

$$\sigma(m_i, m_j) = \exp(\text{sim}(M_i, M_j)/\tau), \quad (1)$$

where M_i and M_j refer to the output molecular representations of m and s from a CLM, respectively. The τ is the temperature parameter for scaling. We employ the NT-Xent loss function (Chen et al., 2020) to maximize agreement between positive pairs while minimizing agreement between negative pairs. Instead of explicitly sampling negative examples, we treat the other $2(N-1)$ molecules in a batch as negative examples. Note that we project the output molecular representations at the <bos> token from CLMs to match the dimensions of the extracted representations. Given a batch of $\{m_1, m_2, m_3, ..., m_{2N}\}$, our loss function for a molecule m_i is defined as follows:

$$\mathcal{L}_{CL}(m_i) = -\log \frac{\sigma(m_i, m_s)}{\sum_{k=1}^{2N-1} \sigma(m_i, m_k)}, \quad (2)$$

where m_s is the similar molecule of m_i in a batch.

3.2 Molecular Substructure Prediction

To further enhance the structural understanding of CLMs, we train the model to predict molecular substructures hashed in fingerprint vectors. We

[2] https://www.rdkit.org

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	<u>69.6</u>	87.8	<u>81.6</u>	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**	<u>84.7</u>	65.6	<u>94.8</u>	80.6	<u>82.2</u>	<u>88.2</u>	66.9	<u>82.1</u>
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)	<u>92.9</u>	83.4	**72.8**	**95.0**	81.3	**82.9**	**89.1**	<u>68.8</u>	**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 <u>underlined</u>.

first train the model to predict these fingerprints directly, to detect the presence of substructures, thereby improving the model's understanding of the structural information of molecules. We employ a Binary Cross Entropy (BCE) loss. Then, our final loss function for a molecule m_i is formulated as follows:

$$\mathcal{L}(m_i) = \mathcal{L}_{BCE}(m_i, f_i) + \lambda \mathcal{L}_{CL}(m_i), \quad (3)$$

where f_i is a fingerprint vector of a molecule m_i and λ is a non-negative hyper-parameter for balancing the objective functions. To ensure accuracy in learning, contrastive learning is omitted for molecules that are not unique, specifically when there are more than two similar molecules within a batch for a particular molecule.

3.3 Fine-tuning of CLMs

After the Moleco training, we add a prediction head to a CLM and fine-tune the model on a target molecular property prediction task. The objective function of the task is as follows:

$$L_{FT}(x) = -\log P(y|x), \quad (4)$$

where $P(\cdot)$ is the prediction of a CLM, x is an input molecule, and y is its prediction label. While Moleco is agnostic to the CLM architecture, in our experiments, we mainly use MoLFormer-XL (Ross et al., 2022) with its pre-trained parameters as our model architecture. MoLFormer-XL is a transformer-based CLM using linear attention with rotary embeddings, modeling molecules in a bi-directional manner.

4 Experiments

4.1 Experimental Settings

Datasets. To evaluate the molecular property prediction ability of CLMs, we conduct experiments on eight classification and four regression tasks from the MoleculeNet benchmark (Wu et al., 2018). For evaluation metrics, we report AUC-ROC for classification, MAE for QM9, and RMSE for remaining regression tasks.

Training Setup. We train models with Moleco on each dataset of the downstream tasks before fine-tuning them. In our experiments, CLMs are initialized with a publicly released MoLFormer-XL checkpoint. For the fine-tuning, we adhere to the recommended train, validation, and test splits from Wu et al. (2018) and follow the experimental settings established by the baseline (Ross et al.,

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)	<u>0.274</u>	<u>0.315</u>	<u>0.540</u>	<u>0.376</u>
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 <u>underlined</u>.

2022). The hyper-parameter settings used for the experiment are shown in Table 8 in Appendix. All experiments are conducted on two NVIDIA RTX A6000 GPUs and four NVIDIA RTX A5000 GPUs.

Baselines. We compare our models with diverse state-of-the-art baselines in three categories. "3D Conformation" includes methods that utilize the geometry information of molecules. "2D Graph" includes methods that utilize 2D graphs including atoms and bonds. "1D SMILES/SELFIES" includes CLMs that utilize string-based descriptors, which are compatible with our Moleco framework.

4.2 Experimental Results

Main Results. We first compare Moleco with state-of-the-art molecular property prediction methods on MoleculeNet classification tasks. As shown in Table 1, Moleco surpasses the state-of-the-art baseline, MoLFormer-XL, by an average of 1.8%. Notably, Moleco exhibits the best performance on 4 tasks and the seconed-best performance on 2 tasks among the 8 tasks. Moreover, as shown in Table 2, Moleco consistently stands out in three MoleculeNet regression tasks, surpassing the state-of-the-art baseline MoLFormer-XL by an average of 7.3%. These results show that contrastive learning based on structural similarity with Moleco can lead to performance improvements in diverse molecular property prediction tasks.

We further compare Moleco with the baselines on QM9, a benchmark on quantum mechanical properties of molecules, as shown in Table 3. Since the quantum mechanical properties are closely related to geometry information of atoms in molecules, methods with ground-truth geometry information (3D Conformation (GT)) achieve the best performances in our experiments. However, this ground-truth geometry information can be obtained through wet lab experiments or massive calculations, which are unavailable in many real-world scenarios as in the above experiments on molecular property classification and regression tasks. In these contexts, we focus on investigating how effectively chemical models can approximate the quantum mechanical properties without such geometry information, by comparing Moleco with 3D methods using geometry information derived by the RDKit library. Our Moleco provides the most accurate prediction of quantum properties without ground-truth geometry information, exhibiting 17.5% of improvements in average over baselines that estimate geometry information or those without any geometry information, demonstrating its efficacy and wide applicability.

Topological Analysis. Following Ross et al. (2022), we evaluate the encapsulated topological information of Moleco by analyzing the resemblance between molecular structures and the attention matrices. We calculate the cosine similarities between average pooled attention matrices and molecular structures. To facilitate this, we randomly select 3,000 molecules from QM9, PubChem (Kim et al., 2019), and ZINC (Irwin et al., 2012) datasets and extract bond connectivity and 3D distance matri-

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)	$C_v \downarrow$ ($\frac{\text{cal}}{\text{mol·K}}$)	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.

Methods	QM9		PubChem		ZINC	
	Bond	Dist.	Bond	Dist.	Bond	Dist.
MoLFormer-XL	60.99	85.73	45.18	79.68	44.11	77.17
Moleco	**62.27**	**87.44**	**45.76**	**80.67**	**44.31**	**78.89**

Table 4: Evaluation of encapsulated topological information. We use Moleco trained on QM9 dataset.

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

Table 5: Comparisons of Moleco variants. CLS and REG denote an average score on molecular property classification and regression tasks, respectively.

ces using RDKit. The results in Table 4 show the Moleco trained on the QM9 dataset exhibits higher similarities across all datasets than its backbone, indicating that Moleco can effectively enhance the capability of identifying molecular structures.

Analysis on Moleco Variants. To verify the efficacy of our design choice, we analyze diverse variants of Moleco using other fingerprint algorithms, similarity functions, and structural embeddings. We evaluate Moleco variants on eight MoleculeNet classification tasks and three MoleculeNet regression tasks except for QM9. The results are shown in Table 5. We first identify that the best setting of Moleco is using Morgan fingerprints and the cosine similarity function. We identify that using fingerprints derived by other algorithms, such as Torsion fingerprint or RDKit fingerprint, and a different similarity function, such as the Tanimoto similarity, degrades the molecular property prediction performance. In addition, we further examine more complex and sophisticated methods to generate molecular embeddings including structural information by calculating similarities using 3D GeoFormer models (Wang et al., 2023b). Surprisingly, this leads to a significant performance degradation, even underperforming original MoLFormer-XL models. We suspect that CLMs and GNNs may have highly different, incompatible views on molecules, particularly about determining the similarities of molecules. We plan to investigate the detailed reason for the incompatibility and integration methods of both models' representations.

Ablation Study To assess the distinct contributions of Moleco's components to its enhanced performance, we conduct ablation studies on three regression tasks with Moleco, detailed in Table 6. These demonstrate that the integration of the two objective functions offers advantages over employing either method in isolation. Furthermore, using our contrastive learning method alone resulted in performance gains on ESOL and FreeSolv. This finding implies that understanding the relationships among molecules facilitates the effective integration of topological information.

	MCL	MSP	ESOL	FreeSolv	Lipop
Moleco	✓	✓	**0.264**	**0.296**	**0.518**
	✓	-	0.292	0.338	0.529
	-	✓	0.286	0.306	0.536
	-	-	0.274	0.315	0.540

Table 6: Ablation study results. MCL and MSP refer to molecular contrastive learning and molecular substructure prediction, respectively.

	ESOL	FreeSolv	Lipop
MoLFormer-XL	0.274	0.315	0.540
MoLFormer-XL + SimCSE	0.280	0.341	0.538
Moleco	**0.267**	**0.296**	**0.518**

Table 7: Comparison of contrastive learning methods on regression tasks.

Contrastive Method Comparison We analyze the impact of different contrastive learning methods on molecular property prediction by comparing MoLFormer-XL with two methods: SimCSE (Gao et al., 2021) and Moleco. Table 7 presents the results on three regression tasks. MoLFormer-XL combined with SimCSE shows either slight performance degradation or minimal improvement compared to the baseline, indicating that SimCSE's random dropout-based data augmentation technique is less effective in this context. In contrast, Moleco consistently outperforms other methods across all datasets, demonstrating its ability to generate chemically meaningful contrastive pairs that better capture the underlying molecular properties.

5 Conclusion

We have introduced Moleco, a novel contrastive learning framework to enhance the structural understanding of CLMs to improve molecular property prediction. We have trained CLMs to contrast structurally similar and dissimilar molecules, which are identified by using the fingerprint vectors of molecules. We have observed that Moleco outperforms state-of-the-art models on diverse molecular property prediction benchmarks. Furthermore, our in-depth analysis has confirmed that Moleco effectively improves the structural understanding of CLMs, leading to significant performance improvements. Particularly, fingerprints, which are highly simplified embedding methods, have most effectively improved the molecular property prediction of CLMs among diverse design choices. We plan to investigate the applicability of Moleco on multi-modal Transformers and generative CLMs.

Acknowledgements

We gratefully acknowledge the support from POSCO Holdings on this work. This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.RS-2024-00415812 and No.2021R1A2C3010430) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2024-00439328, Karma: Towards Knowledge Augmentation for Complex Reasoning (SW Starlab)).

References

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2022. Chemberta-2: Towards chemical foundation models. *arXiv preprint arXiv:2209.01712*.

Keith T Butler, Daniel W Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. 2018. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555.

Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. 2015. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*.

Connor W. Coley, Regina Barzilay, William H. Green Jr., Tommi S. Jaakkola, and Klavs F. Jensen. 2017. Convolutional embedding of attributed molecular graphs for physical property prediction. *J. Chem. Inf. Model.*, 57(8):1757–1772.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. 2022. Geometry-enhanced molecular representation learning for property prediction. *Nat. Mach. Intell.*, 4(2):127–134.

Shikun Feng, Yuyan Ni, Minghao Li, Yanwen Huang, Zhi-Ming Ma, Wei-Ying Ma, and Yanyan Lan. 2024. Unicorn: A unified contrastive learning approach for multi-view molecular representation learning. *arXiv preprint arXiv:2405.10343*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

David E. Graff, Edward O. Pyzer-Knapp, Kirk E. Jordan, Eugene I. Shakhnovich, and Connor W. Coley. 2023. Evaluating the roughness of structure–property relationships using pretrained molecular representations. *Digital Discovery*, 2:1452–1460.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. 2022. Graphmae: Self-supervised masked graph autoencoders. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Weihua Hu, Bowen Liu, Joseph Gomes, Marinka Zitnik, Percy Liang, Vijay S. Pande, and Jure Leskovec. 2020. Strategies for pre-training graph neural networks. In *8th International Conference on Learning Representations*.

John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. 2012. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.*, 52(7):1757–1768.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan Bolton. 2019. Pubchem 2019 update: improved access to chemical data. *Nucleic Acids Res.*, 47(Database-Issue):D1102–D1109.

Johannes Klicpera, Janek Groß, and Stephan Günnemann. 2020. Directional message passing for molecular graphs. In *8th International Conference on Learning Representations*.

Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C. Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, Rafael F. Lameiro, Dominik Lemm, Alston Lo, Seyed Mohamad Moosavi, José Manuel Nápoles-Duarte, AkshatKumar Nigam, Robert Pollice, Kohulan Rajan, Ulrich Schatzschneider, Philippe Schwaller, Marta Skreta, Berend Smit, Felix Strieth-Kalthoff, Chong Sun, Gary Tom, Guido Falk von Rudorff, Andrew Wang, Andrew D. White, Adamo Young, Rose Yu, and Alán Aspuru-Guzik. 2022. SELFIES and the future of molecular string representations. *Patterns*, 3(10):100588.

Taojie Kuang, Yiming Ren, and Zhixiang Ren. 2024. 3d-mol: A novel contrastive learning framework for molecular property prediction with 3d information. *Pattern Analysis and Applications*, 27(3):71.

Tu Le, V Chandana Epa, Frank R Burden, and David A Winkler. 2012. Quantitative structure–property relationship modeling of diverse materials properties. *Chemical reviews*, 112(5):2889–2919.

Shengchao Liu, Weitao Du, Zhi-Ming Ma, Hongyu Guo, and Jian Tang. 2023a. A group symmetric stochastic differential equation model for molecule multi-modal pretraining. In *Proceedings of the 40th International Conference on Machine Learning*.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2022. Pre-training molecular graph representation with 3d geometry. In *10th International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023b. Molca: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhiyuan Liu, Yaorui Shi, An Zhang, Enzhi Zhang, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. 2023c. Rethinking tokenizer and decoder in masked graph modeling for molecules. *Advances in Neural Information Processing Systems*, 36.

Yvonne C Martin, James L Kofron, and Linda M Traphagen. 2002. Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry*, 45(19):4350–4358.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. 2019. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Thirty-Three AAAI Conference on Artificial Intelligence*, 01.

David Rogers and Mathew Hahn. 2010. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754.

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. 2020. Self-supervised graph transformer on large-scale molecular data. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. 2022. Large-scale chemical language representations capture molecular structure and properties. *Nat. Mac. Intell.*, 4(12):1256–1264.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. 2017. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30.

Eduardo Soares, Emilio Vital Brazil, Karen Fiorela Aquino Gutierrez, Renato F. G. Cerqueira, Daniel P. Sanders, Kristin Schmidt, and Dmitry Yu. Zubarev. 2023. Beyond chemical language: A multimodal approach to enhance molecular property prediction. *arXiv preprint arXiv:2306.14919*.

Hannes Stärk, Dominique Beaini, Gabriele Corso, Prudencio Tossou, Christian Dallago, Stephan Günnemann, and Pietro Lió. 2022. 3d infomax improves gnns for molecular property prediction. In *Proceedings of the 39th International Conference on Machine Learning*.

Shantanu Thakoor, Corentin Tallec, Mohammad Gheshlaghi Azar, Mehdi Azabou, Eva L Dyer, Remi Munos, Petar Veličković, and Michal Valko. 2022. Large-scale representation learning on graphs via bootstrapping. In *10th International Conference on Learning Representations*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*.

Jinxian Wang, Jihong Guan, and Shuigeng Zhou. 2023a. Molecular property prediction by contrastive learning with attention-guided positive sample selection. *Bioinform.*, 39(5).

Yusong Wang, Shaoning Li, Tong Wang, Bin Shao, Nanning Zheng, and Tie-Yan Liu. 2023b. Geometric transformer with interatomic positional encoding. *Advances in Neural Information Processing Systems*, 37.

Yuyang Wang, Rishikesh Magar, Chen Liang, and Amir Barati Farimani. 2022a. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. *J. Chem. Inf. Model.*, 62(11):2713–2725.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022b. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287.

Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. 2022c. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.*, 4(3):279–287.

David Weininger. 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36.

Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. 2018. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z. Li. 2023. Mole-bert: Rethinking pre-training graph neural networks for molecules. In *11th International Conference on Learning Representations*.

Junwei Yang, Kangjie Zheng, Siyu Long, Zaiqing Nie, Ming Zhang, Xinyu Dai, Wei-Yin Ma, and Hao Zhou. 2024. Mol-ae: Auto-encoder based molecular representation learning with 3d cloze test objective. *bioRxiv*, pages 2024–04.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020*.

Qiying Yu, Yudi Zhang, Yuyan Ni, Shikun Feng, Yanyan Lan, Hao Zhou, and Jingjing Liu. 2024. Multimodal molecular pretraining via modality blending. In *12th International Conference on Learning Representations*.

Atakan Yüksel, Erva Ulusoy, Atabey Ünlü, Gamze Deniz, and Tunca Dogan. 2023. Selformer: Molecular representation learning via SELFIES language models. *CoRR*, abs/2304.04662.

Xuan Zang, Xianbing Zhao, and Buzhou Tang. 2023. Hierarchical molecular graph self-supervised learning for property prediction. *Communications Chemistry*, 6(1):34.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. 2023. Uni-mol: A universal 3d molecular representation learning framework. In *11th International Conference on Learning Representations*.

Yanqiao Zhu, Dingshuo Chen, Yuanqi Du, Yingze Wang, Qiang Liu, and Shu Wu. 2022. Featurizations matter: A multiview contrastive learning approach to molecular pretraining. In *ICML 2022 2nd AI for Science Workshop*.

Appendix

In this section, we supplement our main content with additional experiments and analysis. We mainly report the results on three molecular property regression tasks (i.e., ESOL, FreeSolv, Lipophilicity) due to the stability of performances on them and high correlations with the average performance on the other tasks.

	Moleco
Backbone	MoLFormer-XL
# Pram.	46M
Batch Size	{32, 64, 128, 256}
Learning Rate	{1e-5, 2e-5, 3e-5, 4e-5, 5e-5}
λ	{0.1, 0.2, 0.3, 0.4, 0.5}
# Mols	{1, 5, 10, 50}
Epoch	{10, 30, 50, 100}

Table 8: Training hyper-parameters for Moleco.

	Epochs	ESOL	FreeSolv	Lipop
Moleco	100	**0.264**	0.327	0.526
	50	0.276	0.330	0.522
	30	0.277	0.310	0.529
	10	0.267	**0.296**	**0.518**
	0	0.274	0.315	0.540

Table 9: Ablation study of contrastive learning. Results with 0 epoch refer to fine-tuning without Moleco.

	# Mols	ESOL	FreeSolv	Lipop
Moleco	top-50	0.298	0.316	0.533
	top-10	0.275	0.315	0.519
	top-5	**0.264**	**0.296**	**0.518**
	None	0.274	0.315	0.540

Table 10: Evaluation of number of similar molecules (# Mols) for the fingerprint-based contrastive learning. Results with None refer to fine-tuning without Moleco.

	Source	ESOL	FreeSolv	Lipop
Moleco	QM9	0.276	**0.274**	0.526
	ESOL	**0.264**	0.355	0.535
	FreeSolv	0.283	0.296	0.530
	Lipop	0.273	0.351	**0.518**
	None	0.274	0.315	0.540

Table 11: Evaluation of the transfer of topological information. Source refers to the dataset used to train Moleco. Results with None refer to fine-tuning without Moleco.

A Additional Analysis

Tables 9 and 10 show hyper-parameter analysis on Moleco. We evaluate the performances on three regression tasks with diverse numbers of epochs and top similar molecules. We have identified that contrastive learning with top-5 similar molecules as positive examples for 10 epochs is the optimal setting of Moleco in our experiments.

We further evaluate the generalizability of molecular representations obtained by Moleco. By training the Moleco framework on three different regression tasks, we cross-evaluate each model with unseen data. The results in Table 11 often show improved performance across these tasks, especially for Moleco with QM9. This highlights the capability of Moleco to effectively transfer topological information, confirming its wide applicability and robustness in boosting performance across various regression tasks.

B Correlation between Molecular Structure and Property

In Section 4.2, we demonstrate Moleco's ability to capture the structural information by following (Ross et al., 2022). This ability is crucial to property prediction. To investigate the correlation between the ability to capture structural information and the predictive performance, we first construct two groups by randomly sampling 30 molecules from the test set. We then evaluate each group using Moleco, reporting the RMSE and cosine similarity of the attention matrix against ground-truth molecular structures (the Bond matrix and the 3D distance matrix). We term the group with relatively higher similarity as "Group 1". The results presented in Tables 12-14 show that "Group 1", which exhibits higher similarity while showing lower RMSEs compared to "Group 2". These findings imply that a deep understanding of structural information is crucial to property prediction.

We attempt various methods to find the most similar molecule set with effective contrast learning, including cosine similarity, string match, and random match. In this process, we first identify the most similar molecules for each molecule using each measurement and then calculate the Mean Absolute Error (MAE) and Maximum Absolute Error (MaxAE) between the properties of the two molecules to compare the results. Consequently, we observe that higher cosine similarity between two molecules tends to exhibit more similar proper-

ESOL	Bond	Dist.	RMSE
Group 1	**53.35**	**87.57**	**0.256**
Group 2	49.90	85.11	0.374

Table 12: Evaluation of two groups of 30 randomly sampled molecules from the ESOL test set.

FreeSolv	Bond	Dist.	RMSE
Group 1	**59.60**	**88.94**	**0.403**
Group 2	57.71	88.77	0.434

Table 13: Evaluation of two groups of 30 randomly sampled molecules from the FreeSolv test set.

Lipop	Bond	Dist.	RMSE
Group 1	**54.43**	**85.20**	**0.421**
Group 2	48.94	83.29	0.587

Table 14: Evaluation of two groups of 30 randomly sampled molecules from the Lipophilicity test set.

ties. Table 15 illustrates that our similarity measurement often results in the minimal average difference in ground-truth properties (MAE) between the query molecule and its top-1 similar counterpart. Furthermore, our similarity measurement proves to be the most effective even in cases of large differences (MaxAE).

C Analysis on Top-5 Selected Molecules

In this section, we qualitatively analyze the selected top-5 molecules. This analysis was conducted on the QM9 dataset using our proposed method, which focuses on identifying molecules with similar properties. Initially, as seen in Figure 2, the similarity distribution of the top-5 selected molecules shows that over 90% have a similarity score of 0.7 or higher, indicating a high level of consistency in the selection process. Additionally, as shown in Figure 3, the selected molecules indeed share mostly similar substructures, suggesting that our method effectively identifies relevant molecular features. These results indicate that our fingerprint-based similarity measure works effectively.

D Extracting Additional Features

We analyze the additional computatinal costs incurred by the process of extracting the similarity features and identifying similar molecules. Notably, we observed that identifying similar molecules is more time-consuming than the feature extraction process itself. Furthermore, as indicated in Table 16, the identification time required for these operations escalates with the increase in dataset size, potentially hindering the application of the Moleco framework in the pre-training phase for enhancements. This highlights the necessity for more efficient algorithms for identifying similar molecules as a pivotal consideration, aiming to streamline the application of the Moleco framework and optimize pre-training efforts.

E Full results of Moleco Variants

We report the full evaluation results of Moleco variants in Tables 19 and 20.

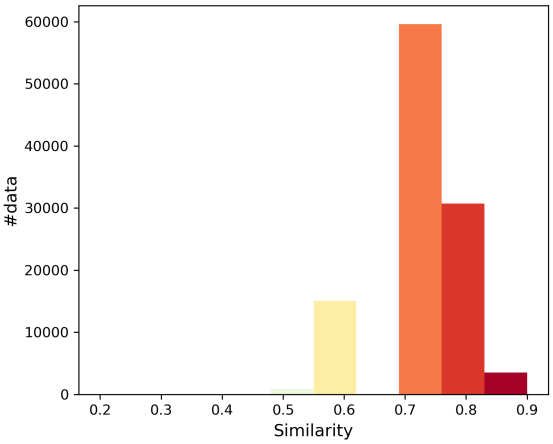

Figure 2: Similarity distribution of the top-5 selected molecules from the QM9 dataset.

	ESOL		FreeSolv		Lipop	
	MAE	MaxAE	MAE	MaxAE	MAE	MaxAE
Cosine similarity	0.71	**3.00**	**1.92**	**11.52**	**0.64**	**2.70**
String match	**0.68**	5.35	2.03	14.27	0.81	3.69
Random match	2.01	7.24	3.84	14.93	1.19	4.40

Table 15: Evaluation of similarity measurement on 50 randomly sampled pairs of top-1 similar molecules and their corresponding queries. We report the Mean Absolute Error (MAE) and Maximum Absolute Error (MaxAE) between the ground-truth properties of molecules. We use the difflib library to calculate the similarity between strings.

	# samples	Extraction time (sec)	Identification time (sec)
FreeSolv	642	< 1	11
ESOL	1,128	< 1	12
SIDER	1,427	< 1	12
ClinTox	1,478	< 1	12
BACE	1,513	1	11
BBBP	2,039	1	12
Lipophilicity	4,200	2	13
Tox21	7,831	3	17
ToxCast	8,577	4	19
HIV	41,127	24	95
MUV	93,087	31	733
QM9	133,885	44	892

Table 16: Time required for extracting ECFP4 fingerprints and identifying similar molecules. We use an NVIDIA A5000 GPU with Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz for this experiment.

	Descriptions	# targets	# examples
BBBP	Blood brain barrier penetration ability	1	2,039
Tox21	Toxicity measurements on 12 targets	12	7,831
ToxCast	Toxicity measurements on 617 targets	617	8,577
Clintox	Toxicity of drugs in clinical trials	2	1,478
MUV	Maximum unbiased validation	17	93,087
HIV	Ability to inhibit HIV replication	1	41,127
BACE	Inhibitors of bindings to human β-secretase 1	1	1,513
SIDER	Side effects on 27 organs	27	1,427

Table 17: Classification tasks from MoleculeNet.

	Descriptions	# targets	# examples
QM9	12 quantum mechanical properties	12	133,885
ESOL	Water solubility of compounds	1	1,128
FreeSolv	Hydration free energy	1	642
Lipophilicity	Solubility in lipids	1	4,200

Table 18: Regression benchmarks from MoleculeNet.

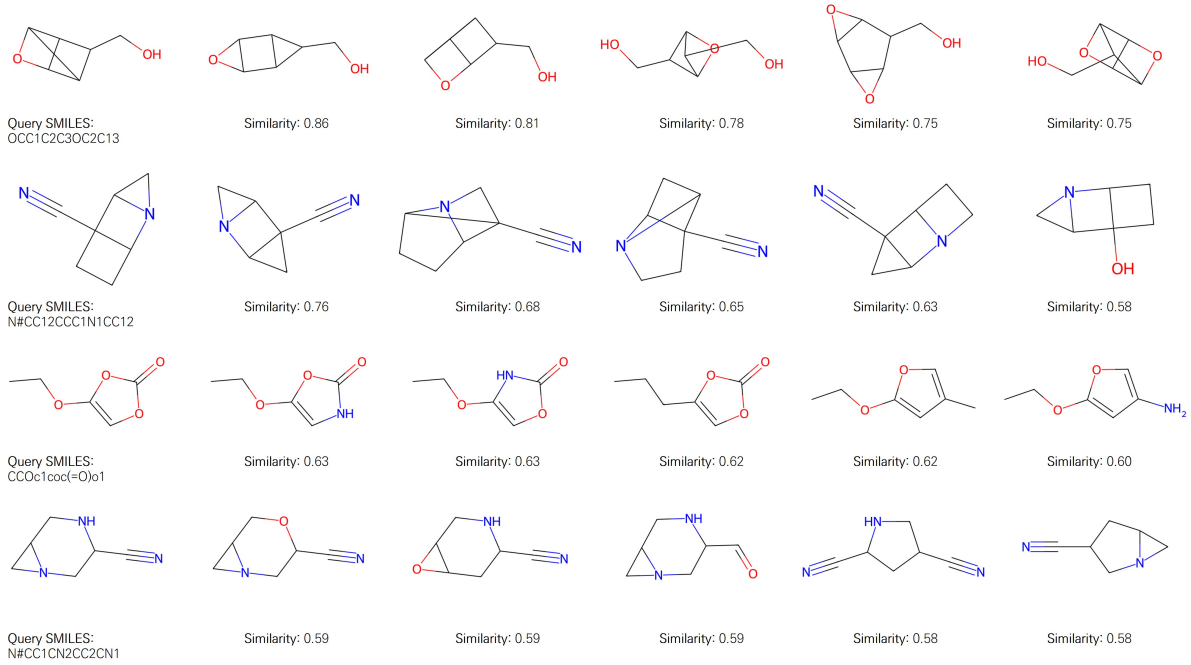

Figure 3: Visualization of the top pairs in the QM9 dataset.

Methods	BBBP ↑	Tox21 ↑	ToxCast ↑	ClinTox ↑	MUV ↑	HIV ↑	BACE ↑	SIDER ↑	Avg. ↑
Moleco-3DGeoFormer-Cosine	92.1	83.5	70.2	93.1	79.6	79.5	83.8	63.1	80.6
Moleco-RDKitFP-Cosine	92.1	**84.5**	70.8	87.3	81.2	80.1	89.1	67.7	81.6
Moleco-TorsionFP-Cosine	91.9	83.8	70.3	92.2	79.3	81.3	89.1	68.3	82.0
Moleco-MorganFP-Tanimoto	**93.1**	84.3	70.9	93.5	**81.6**	77.4	**90.4**	67.3	82.3
Moleco-MorganFP-Cosine	92.9	83.4	**72.8**	**95.0**	81.3	**82.9**	89.1	**68.8**	**83.3**
Moleco w/ ChemBERTa-2	71.4	49.9	50.8	53.5	47.1	74.2	82.8	50.9	60.2

Table 19: Full results of Moleco variants on molecular property classification tasks (ROC-AUC; higher is better)

Methods	ESOL ↓	FreeSolv ↓	Lipophilicity ↓	Avg. ↓
Moleco-3DGeoFormer-Cosine	0.277	0.341	0.519	0.379
Moleco-RDKitFP-Cosine	0.287	0.329	0.524	0.380
Moleco-TorsionFP-Cosine	0.282	0.341	0.527	0.383
Moleco-MorganFP-Tanimoto	0.276	0.329	**0.517**	0.374
Moleco-MorganFP-Cosine	**0.264**	**0.296**	0.518	**0.359**
Moleco w/ ChemBERTa-2	0.811	1.806	0.705	1.107

Table 20: Full results of Moleco variants on molecular property regression tasks (RMSE; lower is better).