MolTRES: Improving Chemical Language Representation Learning for Molecular Property Prediction

Jun-Hyung Park[1] **Yeachan Kim**[2] **Mingyu Lee**[2] **Hyuntae Park**[2] **SangKeun Lee**[2,3]
[1]Division of Language & AI, Hankuk University of Foreign Studies
[2]Department of Artificial Intelligence, Korea University
[3]Department of Computer Science and Engineering, Korea University
jhp@hufs.ac.kr {yeachan, decon9201, pht0639, yalphy}@korea.ac.kr

Abstract

Chemical representation learning has gained increasing interest due to the limited availability of supervised data in fields such as drug and materials design. This interest particularly extends to chemical language representation learning, which involves pre-training Transformers on SMILES sequences – textual descriptors of molecules. Despite its success in molecular property prediction, current practices often lead to overfitting and limited scalability due to early convergence. In this paper, we introduce a novel chemical language representation learning framework, called MolTRES, to address these issues. MolTRES incorporates generator-discriminator training, allowing the model to learn from more challenging examples that require structural understanding. In addition, we enrich molecular representations by transferring knowledge from scientific literature by integrating external materials embedding. Experimental results show that our models outperform existing state-of-the-art models on popular molecular property prediction tasks.

github.com/irishev/MolTRES

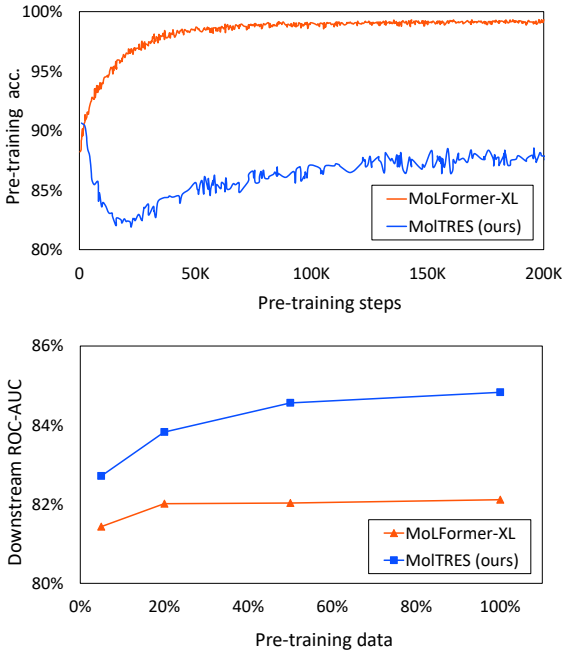

Figure 1: Existing pre-training methods for chemical language representation learning already converge at their early stage without seeing the entire data. Consequently, MoLFormer (Ross et al., 2022), a state-of-the-art chemical language representation learning method, exhibits limited scalability in terms of data size.

1 Introduction

Deep neural networks (DNNs) have emerged as a compelling, computationally efficient approach for predicting molecular properties, with significant implications in material engineering and drug discovery. By training DNNs on molecule data to predict the properties in a supervised manner or to reconstruct molecules in an unsupervised manner, these networks can significantly reduce the costs of traditional methods, which typically require chemical experts and wet-lab experiments. Moreover, DNN-based molecular prediction has gained increasing popularity due to the generalization capacity of DNNs. This allows for the application of a single (pre-)trained model across various tasks, reducing the need for task-specific modeling.

Inspired by recent advances in pre-trained language models in the field of natural language processing (NLP), several chemical language representation learning methods based on Transformers (Wang et al., 2019; Chithrananda et al., 2020) have been proposed. These methods typically employ self-supervised tasks on SMILES (Simplified Molecular-Input Line Entry System) sequences of molecules, analogous to the masked language modeling (MLM) commonly used in BERT (Devlin et al., 2019). Since modern Transformers are designed to scale to massive NLP corpora (Vaswani et al., 2017), they offer practical advantages in terms of efficiency and throughput. This enables the models to leverage massive amounts

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 14241–14254
November 12-16, 2024

of SMILES sequences to learn universal representations for molecules, leading to performance improvements in a wide range of molecular property prediction tasks (Ross et al., 2022). However, as these models typically follow settings designed for natural language modeling, the optimal pre-training settings for chemical language representation learning remain underexplored.

Through extensive investigation into the pre-training of SMILES Transformers, we have discovered that the current pre-training task, MLM on SMILES sequences using a random masking strategy, is not effective for learning informative molecular representations. We have empirically observed that this task can be easily solved using surface patterns, leading to overfitting and limited scalability, as shown in Figure 1. This may be attributed to two inherent properties of SMILES. First, existing large-scale molecule datasets exhibit unbalanced atom distributions (He et al., 2023a). For example, in ZINC (Irwin et al., 2012), a representative dataset containing billions of molecules, carbon (C), nitrogen (N), and oxygen (O) comprise 95% of the tokens in total SMILES sequences. Second, the SMILES grammar contains many superficial patterns, such as numbers representing ring structures that always appear twice. These patterns allow the model to predict original tokens by using simple rules, without learning the underlying chemical information. Furthermore, unlike natural language, which is fundamentally grounded in concepts and possesses general expressivity across various problem-solving scenarios, SMILES is designed solely to express molecular structure and does not directly represent molecular properties. Thus, the current pre-training task likely provides a limited notion of molecular properties.

In this paper, we propose a novel framework for pre-training SMILES transformers, called MolTRES (**Mol**ecular **TR**ansformer with **E**nhanced **S**elf-supervised learning), to address the aforementioned issues. Our framework focuses on two key objectives: (1) increasing the difficulty of the pre-training task, and (2) incorporating external knowledge related to molecular properties into model representations. To achieve these goals, we first present a novel self-supervised learning pipeline, coined DynaMol, based on generator-discriminator training (Clark et al., 2020). This method trains a model to distinguish real SMILES tokens from synthetically generated replacements, jointly used with substructure-level masking. It facilitates to significantly increase the masking ratio for more challenging training examples that require an understanding of molecular structure, while minimizing discrepancy caused by mask tokens. In addition, we enhance model representations by integrating mat2vec word representations (Tshitoyan et al., 2019) trained on massive scientific literature. This integration helps to directly embody molecular properties in the learned representations.

To demonstrate the effectiveness of MolTRES, we conduct extensive experiments and ablation studies on diverse molecular property prediction tasks. We evaluate MolTRES on eight classification and four regression tasks from MoleculeNet (Wu et al., 2018), covering quantum mechanical, physical, biophysical, and physiological properties of chemicals. Our results indicate that MolTRES outperforms state-of-the-art baselines across most tasks, including 1D sequence-, 2D graph-, and 3D geometry-based chemical models. Furthermore, we observe that MolTRES outperforms state-of-the-art models on seven polymer property prediction tasks, showing its generalizability to different chemical tasks. Further analysis shows that MolTRES significantly improves the capabilities of chemical language representation learning by addressing the limitations of existing approaches. Our contributions are summarized as follows:

- We propose MolTRES, a novel framework to pre-train SMILES Transformers based on generator-discriminator training and external knowledge transfer.
- We present a novel architecture for SMILES transformers efficiently integrated with word representations trained on scientific literature.
- Experimental results demonstrate that MolTRES establishes state-of-the-art results over a wide range of molecular property prediction tasks.

2 Related Work

In recent years, representation learning has prevailed in numerous applications in natural language processing (Devlin et al., 2019; Liu et al., 2019) and computer vision (Dosovitskiy et al., 2021; Bao et al., 2021). This trend has triggered many studies in chemical representation learning. The approaches in this field can be classified into three categories based on molecular descriptors used for pre-training: chemical language representation

learning, chemical graph representation learning, and multi-modal chemical representation learning.

Chemical language representation learning. Chemical language representation learning has adopted pre-training on molecular descriptors represented as strings, such as SMILES and SELFIES. It typically leverages Transformers (Vaswani et al., 2017) to learn molecular descriptors inspired by the recent success of large-scale representation learning in natural language processing. Wang et al. (2019); Chithrananda et al. (2020); Ross et al. (2022) have trained Transformer models on large-scale SMILES sequences. Yüksel et al. (2023) have utilized SELFIES sequences to achieve a better representation space. However, the training strategies for these methods follow the practice of MLM-style training in natural language processing. Since chemical language differs from natural language, current applications of MLM encounter various issues in pre-training. In this work, we propose MolTRES to address these issues and consequently improve molecular property prediction.

Chemical graph representation learning. Researchers in chemical graph representation learning argue that molecules can naturally be represented in 2D or 3D graphs. Thus, they typically leverage graph neural networks (GNNs) or Transformers adapted to graphs. Hu et al. (2020) have introduced a self-supervised task for molecular graphs, called AttrMask. Morris et al. (2019a) have introduced higher-order GNNs for distinguishing non-isomorphic graphs. You et al. (2020) have extended contrastive learning to unstructured graph data. Wang et al. (2022) have proposed a unified GNN pre-training framework that integrates contrastive learning and sub-graph masking. Recent work has focused on modeling 3D graphs, as they provide more vital information for predicting molecular properties compared to 2D graphs. Yang et al. (2024); Zhou et al. (2023) have proposed denoising auto-encoders for directly modeling 3D graphs. However, due to the limited scale of 3D molecular data and its resource-intensive modeling, the applicability of 3D approaches is limited.

Multi-modal chemical representation learning. Recently, several studies have proposed learning chemical representations in a multi-modal manner, typically leveraging both 2D topology and 3D geometry of molecules. Liu et al. (2022); Stärk et al. (2022); Liu et al. (2023a) have introduced a contrastive learning framework that uses 2D graphs and their corresponding 3D conformations as positive views, treating those from different molecules as negative views. Luo et al. (2022) have proposed encoding both 2D and 3D inputs within a single GNN model. Another research direction has involved using both chemical and natural languages (Edwards et al., 2022; Liu et al., 2023b) to enrich molecular representations and facilitate molecule generation using natural language. This research direction is distantly related to our work, and we plan to further explore the multi-modal and generation capabilities of MolTRES.

3 MolTRES: Molecular Transformer with Enhanced Self-supervised Learning

In this section, we detail our framework, MolTRES, which is illustrated in Figure 2. We propose a novel pre-training task, called DynaMol, which incorporates generator-discriminator training with substructure masking to increase the difficulty of a pre-training task. In addition, we integrate word representations that have been trained on scientific literature to enrich information directly related to molecular properties in the representations.

3.1 DynaMol: Dynamic Molecule Modeling with Generator-Discriminator Training

To increase the difficulty of chemical language representation learning, we propose a dynamic molecule modeling scheme based on generator-discriminator training, inspired by replaced token detection proposed in Clark et al. (2020). The proposed scheme involves training two models, namely a generator and a discriminator. The generator is trained to predict original sequences given masked sequences similar to MLM, while the discriminator is trained to identify tokens that have been replaced by the generator. Since the generator transforms masked sequences to more closely resemble original distributions, this training scheme results in less discrepancy between the inputs from pre-training and downstream tasks, and allows for flexible adjustments of the masking ratio (He et al., 2023b). Moreover, as the generator is being trained, it naturally provides increasingly challenging examples to the discriminator due to a closer distribution towards the true one. This scheme is expected to alleviate the issues of early convergence and overfitting commonly observed in existing methods of chemical language representation learning.

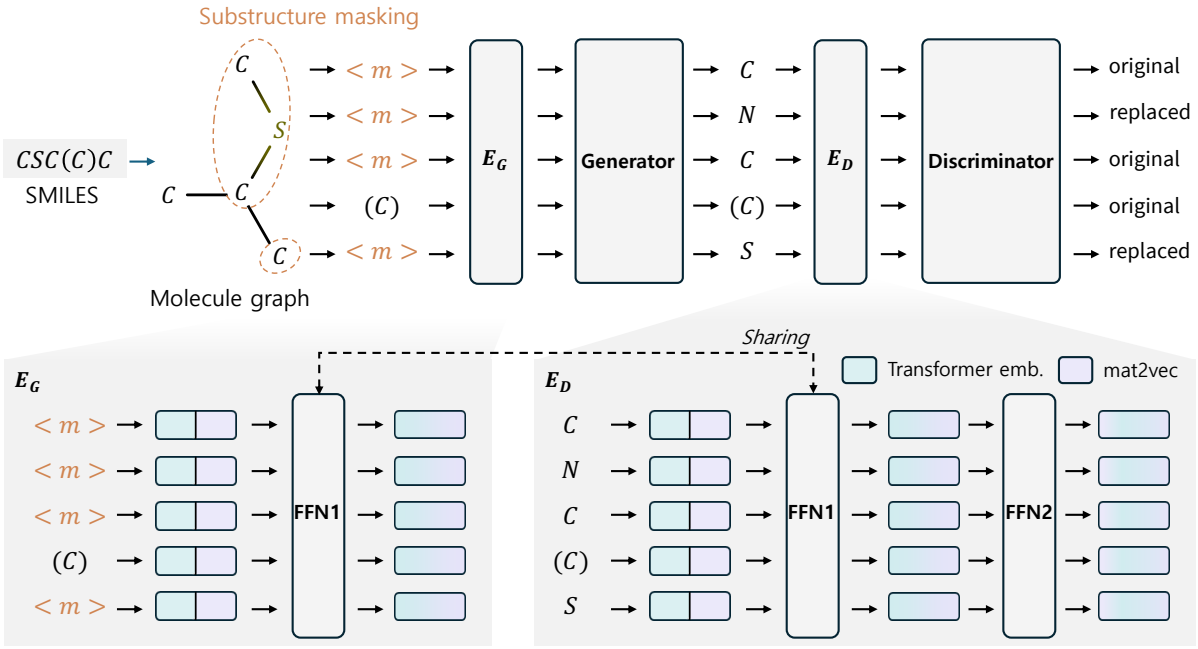

Figure 2: Overview of MolTRES. E_G and E_D represent the embedding layers of the generator and discriminator, respectively. FFNs denote feed-forward networks that linearly project the feature vector of each token. It is noteworthy that the mat2vec embeddings are frozen during pre-training.

Specifically, given a token sequence $\mathbf{X} = \{x_1, x_2, x_3, ..., x_n\}$, we corrupt $\mathbf{X}$ into $\tilde{\mathbf{X}}$ by partly masking tokens. Then, the generator G with parameters θ_G is trained to reconstruct the sequence $\mathbf{X}$. The loss of G for each example is formulated as follows:

$$\mathcal{L}_G = -\sum_{i \in \mathcal{M}} \log p(x_i | \tilde{\mathbf{X}}; \theta_G), \tag{1}$$

where $\mathcal{M}$ represents the set of masked token positions. Similar to typical chemical language representation learning methods, each masked token is substituted with a special mask token in 80% of cases, a random token in 10% of cases, and the original token in the remaining 10% cases, since we have observed that changing the substitution strategy barely affects the performance.

The input sequence for the discriminator is constructed by replacing the masked tokens in $\tilde{\mathbf{X}}$ with new tokens, sampled from the generator's probability distribution, as follows:

$$\tilde{\mathbf{X}}_D = \begin{cases} \tilde{x}_i \backsim p(x_i | \tilde{\mathbf{X}}; \theta_G) & \text{if } i \in \mathcal{M} \\ x_i & \text{otherwise.} \end{cases} \tag{2}$$

The discriminator is trained to distinguish whether each token in the generated input sequence $\tilde{\mathbf{X}}_D$ is original or has been replaced. The loss for the discriminator is formulated as follows:

$$\mathcal{L}_D = -\sum_{i=1}^{n} \log p(z_i | \tilde{\mathbf{X}}_D; \theta_D), \tag{3}$$

where z_i is a binary label that indicates whether the i-th input token is original or has been replaced. Finally, the generator G and discriminator D are jointly optimized with multiple objectives, expressed as $\mathcal{L} = \mathcal{L}_G + \lambda \mathcal{L}_D$, where λ is a predefined balancing parameter for the discriminator loss. In this work, λ is set to 10.

In addition, we carefully design three rules to mask SMILES at multiple substructure-level granularities, thereby preventing models from predicting the correct answer by exploiting superficial patterns in the SMILES grammar. (1) We mask all special tokens that represent structural information, such as numbers for cycles. (2) We then mask spans of SMILES that composes certain substructures, such as substituents, bridges, or groups of sequential atoms, until the ratio of masked tokens does not exceed the pre-defined target masking ratio. Note that these substructures are identified by segmenting SMILES strings based on brackets in our method. (3) We mask random atomic SMILES tokens to achieve the target masking ratio. Notably, we use 65% of the target masking ratio for pre-training.

3.2 Knowledge Transfer from Scientific Literature using mat2vec

While modeling SMILES helps models understand molecular structure and connectivity, SMILES itself lacks explicit information about molecular properties. Scientific literature, which is similarly represented in a textual form, provides a more flexible and rich source of external information. It comprehensively involves information about molecular properties derived from wet laboratory experiments and computational methods. Therefore, we enrich the representations of SMILES Transformers by integrating information from scientific literature.

Despite the many possible design choices available, we opt to leverage mat2vec (Tshitoyan et al., 2019), a straightforward embedding model trained on extensive scientific literature, for integration into Transformer's embedding vectors. We prioritize the efficiency in terms of memory footprints and computations in our integration procedures, essential for dealing with large-scale pre-training. Given an input sequence $\mathbf{X} = \{x_1, ..., x_n\}$, we obtain embedding vectors for every token from the Transformer's embedding layer, denoted as $\mathbf{E}^t = \{e_1^t, ..., e_n^t\}$. Using a mapping function $I(\cdot)$, we assign each token to corresponding mat2vec embedding vectors, denoted as $\mathbf{E}^m = \{e_1^m, ..., e_n^m\}$ s.t. $e_k^m = \sum_{z \in I(x_k)} \text{mat2vec}(z)$. We then combine $\mathbf{E}^t$ and $\mathbf{E}^m$ using a linear projection layer $F_1(\cdot)$. The set of embedding vectors for the generator V_G is generated as follows:

$$\mathbf{V}_G = \{F_1(e_1^t \circ e_1^m), ..., F_1(e_n^t \circ e_n^m)\}, \quad (4)$$

where $\circ$ denotes the concatenation operation. In a similar manner, the set of embedding vectors for the discriminator V_D is generated from the tokens reconstructed by the generator as follows:

$$\begin{aligned} \mathbf{V} &= \{F_1(\tilde{e}_1^t \circ \tilde{e}_1^m), ..., F_1(\tilde{e}_n^t \circ \tilde{e}_n^m)\} \\ \mathbf{V}_D &= \{F_2(\sigma(v_1)), ..., F_2(\sigma(v_n))\} \\ &\text{s.t. } v_1, ..., v_n \in V, \end{aligned} \quad (5)$$

where $v_1, ..., v_n \in V$ and $\sigma(\cdot)$ is an activation function, which is the gelu function in this work.

For the integration, we manually design a mapping function $I(\cdot)$ using human prior knowledge to address the vocabulary mismatch between SMILES tokens and mat2vec words. We utilize a thesaurus carefully constructed by domain experts, chosen for its superior computational efficiency and stability compared to learning-based approaches. For example, the thesaurus maps "[cH+]" in the Transformer's vocabulary to "methylidyne", "ion", and "cation" in the mat2vec vocabulary. Based on this thesaurus, we pre-calculate embedding vectors for 2,696 tokens in the Transformer vocabulary before pre-training. To prevent catastrophic forgetting of mat2vec knowledge, we freeze these pre-calculated embedding vectors during pre-training. During fine-tuning, these embedding vectors are trainable to adapt the knowledge for each downstream task.

4 Experiment

4.1 Experimental Setup

Pre-training. We collect 118 million molecules from PubChem[1] and 1.9 billion molecules from ZINC[2]. We pre-train two MolTRES models, a base model (MolTRES) and a smaller model (MolTRES-small). For pre-processing, we extract the canonicalized format of SMILES for every molecule using RDKit[3]. We construct the vocabulary with 2,691 unique tokens plus five special tokens ("<bos>", "<eos>", "<pad>", "<mask>", and "<unk>") after tokenizing all the extracted SMILES sequences. For tokenization, we use the maximum sequence length of 512. The weights of our models are initialized over the normal distribution with a standard deviation of 0.02. Pre-training is performed using an AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$), where the maximum learning rate and weight decay are set to 3e-4 and 0.01, respectively. We use the cosine annealing for learning rate scheduling with 1,000 warmup steps. We train our models for 200,000 steps with a batch size of 25,600 and use the final models in evaluation. The pre-training time of MolTRES is approximately 15 days using 4 NVIDIA RTX A6000 GPUs.

Evaluation We evaluate our models and baselines on eight classification tasks and four regression tasks from the MoleculeNet benchmark (Wu et al., 2018). We use the scaffold splitting (80% / 10% / 10% for train / validation / test) for all the tasks except for QM9, in which the random split (80% / 10% / 10% for train / validation / test) with thermochemical energy pre-calculation is used following Liu et al. (2023a). We further evaluate our models on seven polymer property prediction tasks with the random split strategy. The statistics of evaluation benchmarks are shown in Ap-

[1] https://pubchem.ncbi.nlm.nih.gov/
[2] https://zinc.docking.org/
[3] 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	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., 2022)	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									
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	-
MolTRES-small (ours)	95.0	83.4	64.8	94.0	80.0	81.7	87.7	68.3	81.9
MolTRES (ours)	**96.1**	**85.3**	**70.1**	**96.7**	**84.9**	**84.2**	**91.7**	69.8	**84.8**

Table 1: Evaluation results on MoleculeNet classification tasks. We report ROC-AUC scores (higher is better) under scaffold splitting. The best and second-best results are in **bold** and underlined.

pendix. For evaluation of our models, we extract the output representations from model's final transformer block corresponding to the first input token ("<bos>") as the molecule representations. We use a 2-layer MLP with the same hidden size and gelu activation as a classifier, whose weights are initialized over the normal distribution with a standard deviation of 0.02. We use the augmentation of random SMILES reconstruction for all the tasks. We fine-tune the models for 500 epochs using an AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.99$) with a weight decay of 0.01. For each task, we empirically choose the batch size $\in \{16, 32, 64, 128\}$ and learning rate $\in$ {2e-5, 3e-5, 5e-5, 1e-4}. We report Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) scores for the classification tasks, Mean Absolute Error (MAE) scores for QM9, and Root Mean Square Error (RMSE) scores for the remaining regression tasks. We report the average test score on five different splits using the models that achieve the best validation scores.

Model Architecture. The model architecture of the generator and discriminator is a Transformer with linear attention and rotary position embeddings, following Ross et al. (2022). The discriminator of MolTRES has 12 layers, 768 hidden dimensions, and 12 attention heads, while that of MolTRES-small has 6 layers, 768 hidden dimensions, and 12 attention heads. The generators have half the number of layers in their corresponding discriminator, while the other settings are consistent. It is noteworthy that the generator is only used for pre-training, and the discriminator is fine-tuned and evaluated in all the downstream tasks. The generator and discriminator share their embeddings as in Clark et al. (2020).

Baselines. We compare our models with diverse state-of-the-art methods in molecular property prediction, categorized as follows:

- **3D Conformation:** This category includes methods that utilize 3D conformation from the geometry information of molecules and may incorporate other modalities.
- **2D Graph:** This category includes methods that utilize 2D graph with atoms and bonds, and may also combine 1D SMILES.
- **1D SMILES/SELFIES:** This category includes methods that utilize SMILES or SELFIES sequences of molecules.

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., 2022)	1.110	2.200	0.650	1.320
SimSGT (Liu et al., 2023c)	0.917	-	0.695	-
1D SMILES/SELFIES				
MoLFormer-Base (Ross et al., 2022)	0.280	0.260	0.649	0.396
MoLFormer-XL (Ross et al., 2022)	<u>0.279</u>	<u>0.231</u>	<u>0.530</u>	<u>0.347</u>
SELFormer (Yüksel et al., 2023)	0.682	2.797	0.735	1.405
MolTRES-small (ours)	0.280	0.250	0.594	0.375
MolTRES (ours)	**0.274**	**0.229**	**0.504**	**0.336**

Table 2: Evaluation results on MoleculeNet regression tasks. We report RMSE scores (lower is better) under scaffold splitting. The best and second-best results are in **bold** and <u>underlined</u>.

Methods	Egc↓ (eV)	Egb↓ (eV)	Eea↓ (eV)	Ei↓ (eV)	Xc↓ (%)	EPS↓	Eat↓ (eV atom^{-1})	Avg.↓
ChemBERTa (Chithrananda et al., 2020)	0.539	0.664	0.350	0.485	18.711	0.603	0.219	3.082
polyBERT (Kuenneth and Ramprasad, 2023)	0.553	0.759	0.363	0.526	18.437	0.618	0.172	3.061
Transpolymer (Xu et al., 2023)	**0.453**	<u>0.576</u>	**0.326**	<u>0.397</u>	<u>17.740</u>	<u>0.547</u>	<u>0.147</u>	<u>2.884</u>
MolTRES (ours)	<u>0.480</u>	**0.496**	<u>0.339</u>	**0.342**	**15.471**	**0.472**	**0.029**	**2.518**

Table 3: Evaluation results on polymer property prediction tasks (Kuenneth et al., 2021). We report RMSE scores (lower is better) with the random splitting. The best and second-best results are in **bold** and <u>underlined</u>.

4.2 Main Results

We first compare MolTRES with state-of-the-art molecular property prediction methods on MoleculeNet classification tasks. As shown in Table 1, MolTRES surpasses the best baseline, MoLFormer-XL, by an average of 2.7%. In addition, MolTRES-small also shows a competitive performance compared to the baselines. Notably, MolTRES significantly outperforms baseline methods using 3D conformation and 2D graph. This confirms the strength of pre-training with billion-scale SMILES sequences, compared to pre-training with hundreds of millions of conformation or graph examples. MolTRES exhibits state-of-the-art performance on 7 of the 8 tasks. Although MolTRES achieves the second-best results after SELFormer on the SIDER task, it outperforms SELFormer by up to 20% on the others, affirming the superiority of MolTRES.

Moreover, as shown in Table 2, MolTRES consistently stands out in three MoleculeNet regression tasks, surpassing the state-of-the-art method MoLFormer-XL by an average of 3.3%. Moreover, MolTRES-small achieves better performance than MoLFormer-Base, which contains a commensurate number of parameters, by an average of 5.6%. The superior performance of SMILES-based methods is still observed, as they achieve significantly smaller errors compared to other baseline methods. This performance gap further verifies the efficacy of large-scale pre-training on SMILES.

We present the performance of MolTRES on polymer property prediction tasks (Kuenneth et al., 2021) in comparison to other SMILES-based models, shown in Table 3. These tasks involve very large molecules, which can be represented using P-SMILES grammar that is compatible with MolTRES. In our results, although Transpolymer and PolyBERT are specifically designed for polymer property prediction, MolTRES, without any modifications, exhibits superior performance on average. These results verify the generalizablity of MolTRES to a wide range of chemical tasks based on SMILES-like grammars. Moreover, our approach is adaptable to other sequence-based chemical tasks, such as protein property prediction using amino acid sequences, opening up many promising

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., 2019b)	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., 2019b)	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
MolTRES-small (ours)	0.326	0.295	0.066	0.067	0.085	16.32	0.009	0.133	0.185	0.155	0.164	0.137	1.495
MolTRES (ours)	**0.315**	**0.237**	**0.054**	**0.057**	**0.077**	**14.60**	0.007	0.061	0.071	0.068	0.057	0.121	**1.310**

Table 4: Evaluation results on QM9 tasks. We report MAE scores (lower is better) following the data splitting used in Liu et al. (2023a). The best and second-best results are in **bold** and underlined. It is important to note that the "3D Conformation (GT)" results utilize ground-truth geometry information, which incurs non-trivial costs to obtain. For a fair comparison, we also evaluate the performance of 3D models using the geometry information approximated by RDKit, denoted as "3D Conformation (RDKit)", considering scenarios where ground-truth geometry is unavailable.

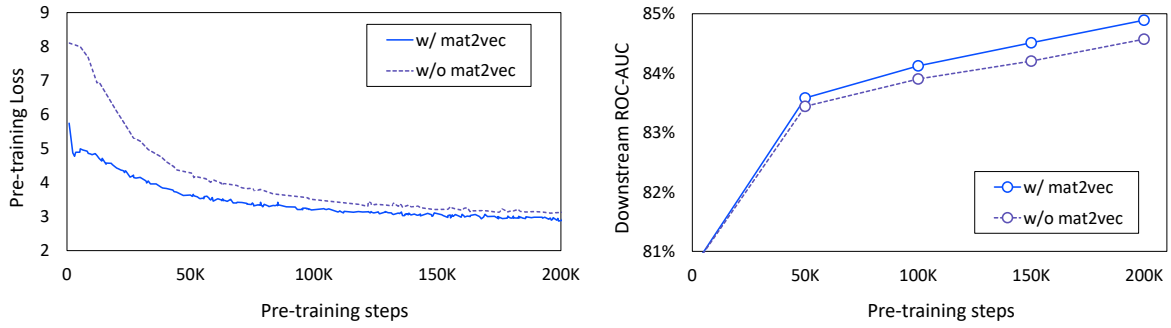

Figure 3: Training curves of MolTRES with mat2vec embeddings (the solid line) and without mat2vec embeddings (the dashed line). The left shows the pre-training loss curves, while the right shows the average ROC-AUC scores.

applications to be investigated.

We further compare MolTRES with the baselines on QM9, as shown in Table 4. Since quantum properties are strongly correlated with geometry information, baselines using ground-truth geometry information (3D Conformation (GT)) show the best results among baselines. However, obtaining this geometry information involves non-trivial costs and may not be available in many real-world scenarios. In these contexts, our MolTRES models provide the most accurate approximation by only using SMILES, compared to baselines that estimate geometry information from RDKit or those without any geometry information, demonstrating its efficacy and applicability. In addition, MolTRES-small model also outperforms all the baselines, showing its efficiency in approximation scenarios.

4.3 Analysis

To better understand the improvements from MolTRES, we report a series of analysis on MoleculeNet classification tasks. The results on regression tasks are provided in Appendix.

Effect of mat2vec embedding. We analyze the effect of the mat2vec embeddings on the pre-training of MolTRES. As described in Figure 3, mat2vec enables faster convergence, attributed to the rich features provided by mat2vec that are beneficial for structure modeling. Additionally, when fully trained, MolTRES with mat2vec achieves lower training losses and enhanced performance in MoleculeNet classification tasks. This validates the effectiveness of integrating mat2vec embeddings.

DynaMol	mat2vec	BBBP ↑	Tox21 ↑	ToxCast ↑	ClinTox ↑	MUV ↑	HIV ↑	BACE ↑	SIDER ↑
✓	✓	**95.2**	**86.1**	**71.9**	**93.9**	**82.7**	**81.1**	**93.7**	**69.1**
✓		**95.2**	85.9	71.2	93.7	81.5	78.5	93.3	68.5
	✓	92.5	84.7	67.0	87.3	80.2	79.0	93.1	66.4
		92.1	84.4	66.1	86.9	78.8	77.1	92.4	64.8

Table 5: Performance on validation sets of MoleculeNet classification tasks with variants of MolTRES.

Masking	BBBP ↑	Tox21 ↑	ToxCast ↑	ClinTox ↑	MUV ↑	HIV ↑	BACE ↑	SIDER ↑
Random	**95.4**	**86.4**	68.3	93.8	82.3	80.7	93.2	68.1
Ours	95.2	86.1	**71.9**	**93.9**	**82.7**	**81.1**	**93.7**	**69.1**

Table 6: Performance on validation sets of MoleculeNet classification tasks with different masking strategies.

Model	BBBP ↑	Tox21 ↑	ToxCast ↑	ClinTox ↑	MUV ↑	HIV ↑	BACE ↑	SIDER ↑
MolTRES gen	93.3	83.3	63.1	91.1	77.9	80.7	87.3	65.8
MolTRES-small disc	**95.0**	**83.4**	**64.8**	**94.0**	**80.0**	**81.7**	**87.7**	**68.3**

Table 7: Performance of the MolTRES generator and discriminator on validation sets of MoleculeNet classification tasks. Note that we use the discriminator of MolTRES-small, which has the same size of the generator of MolTRES.

Ablation Study. To assess the distinct contributions of MolTRES's components to its enhanced performance, we conduct ablation studies using variants of MolTRES as detailed in Table 5. The results demonstrate that both the DynaMol and mat2vec integration contribute to performance improvements. While DynaMol shows consistent, significant performance improvements over the tasks, mat2vec integration particularly exhibits considerable improvements on five tasks (ToxCast, MUV, HIV, BACE, and SIDER). Moreover, when used jointly, they offer complementary advantages over employing either method in isolation. This result underscores the effectiveness of each component in MolTRES in addressing the issues in existing chemical language representation learning, leading to notable performance improvements.

Masking strategy. To evaluate the performance improvements from our substructure masking method, we compare it with a random masking strategy widely used in typical chemical language representation methods, shown in Table 6. The results demonstrate the efficacy of our masking strategy, outperforming random masking on average. We observe that our masking strategy effectively increases the loss value in identifying the original SMILES sequence, addressing the difficulty issue in typical pre-training methods. Furthermore, it may help models identify some functional groups in molecules that causes their characteristic properties, which are similarly observed in BERT with whole-word and PMI masking (Levine et al., 2021) in the field of natural language processing.

Generator vs. Discriminator. Since our pre-training framework involves two models, namely a generator and a discriminator, we compare their performances on molecular property prediction tasks, shown in Table 7. We observe that a generator model performs significantly worse than a same-size discriminator model. In molecular property prediction tasks, applying generation models directly on understanding tasks appears harmful, consistent with the results in natural language processing work (Clark et al., 2020).

5 Conclusion

In this work, we have proposed a novel chemical language representation learning framework, MolTRES. We have identified critical, previously unaddressed issues in existing methods for chemical language representation learning, specifically early convergence and overfitting, and presented two methods, generator-discriminator training with substructure masking and knowledge transfer from scientific literature based on mat2vec. Our experimental results validate the superiority of our framework over existing chemical models across a wide range of molecular property prediction tasks, showing that MolTRES uniquely enables a robust training of chemical language models and provides substantial improvements in performances.

Limitations

While we have demonstrated that MolTRES effectively improves molecular property prediction by addressing issues in existing chemical language representation learning methods, some limitations open promising avenues for future research. First, several components in MolTRES, such as its masking strategy or knowledge transfer method, were chosen empirically in terms of efficiency, and therefore may have room for performance improvements through theoretical or learning-based approaches. Second, we evaluated a few architectural settings of MolTRES corresponding to those of MoLFormer-XL for comparison. Future evaluations could explore more diverse settings of MolTRES to accommodate various scenarios, including resource-limited or scalable environments. Finally, a popular application of SMILES Transformers is in molecule generation. We plan to investigate the extension of MolTRES on the pre-training of generative Transformers for this purpose.

Acknowledgements

This work was 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), No.RS-2024-00457882, AI Research Hub Project, and No.RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University)).

References

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. 2020. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *CoRR*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations*.

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*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *The Nineth International Conference on Learning Representations*.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between molecules and natural language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413.

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*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023b. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Zhenyu Hou, Xiao Liu, Yukuo Cen, Yuxiao Dong, Hongxia Yang, Chunjie Wang, and Jie Tang. 2022. Graphmae: Self-supervised masked graph autoencoders. In *The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 594–604.

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.

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

Christopher Kuenneth, Arunkumar Chitteth Rajan, Huan Tran, Lihua Chen, Chiho Kim, and Rampi Ramprasad. 2021. Polymer informatics with multi-task learning. *Patterns*, 2(4).

Christopher Kuenneth and Rampi Ramprasad. 2023. polybert: a chemical language model to enable fully machine-driven ultrafast polymer informatics. *Nature Communications*, 14(1):4099.

Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. 2021. Pmi-masking: Principled masking of correlated spans. In *The Tenth International Conference on Learning Representations*.

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 *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 21497–21526. PMLR.

Shengchao Liu, Hanchen Wang, Weiyang Liu, Joan Lasenby, Hongyu Guo, and Jian Tang. 2022. Pre-training molecular graph representation with 3d geometry. In *The Tenth 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. *CoRR*, abs/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 *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.

Shengjie Luo, Tianlang Chen, Yixian Xu, Shuxin Zheng, Tie-Yan Liu, Liwei Wang, and Di He. 2022. One transformer can understand both 2d & 3d molecular data. In *The Eleventh International Conference on Learning Representations*.

Christopher Morris, Martin Ritzert, Matthias Fey, William L. Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. 2019a. Weisfeiler and leman go neural: Higher-order graph neural networks. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 4602–4609.

Christopher Morris, Martin Ritzert, Matthias Fey, William L Hamilton, Jan Eric Lenssen, Gaurav Rattan, and Martin Grohe. 2019b. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4602–4609.

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. *Advances in Neural Information Processing Systems*, 33.

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.*, 4:1256–1264.

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.

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 *International Conference on Machine Learning*, volume 162, pages 20479–20502.

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 *The Tenth International Conference on Learning Representations*.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nat.*, 571(7763):95–98.

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*, volume 30, pages 5998–6008.

Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. 2019. SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 429–436.

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

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 *The Eleventh International Conference on Learning Representations*.

Changwen Xu, Yuyang Wang, and Amir Barati Farimani. 2023. Transpolymer: a transformer-based language model for polymer property predictions. *npj Computational Materials*, 9(1):64.

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*, 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*, volume 33.

Qiying Yu, Yudi Zhang, Yuyan Ni, Shikun Feng, Yanyan Lan, Hao Zhou, and Jingjing Liu. 2024. Multimodal molecular pretraining via modality blending. In *The Twelfth 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.

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 *The Eleventh International Conference on Learning Representations*.

DynaMol	mat2vec	ESOL ↓	FreeSolv ↓	Lipo ↓
✓	✓	**0.294**	0.209	0.515
✓		0.298	**0.208**	0.517
	✓	0.296	0.209	0.529
		0.301	0.212	0.540

Table 8: Performance on validation sets of MoleculeNet regression tasks with variants of MolTRES.

Masking	ESOL ↓	FreeSolv ↓	Lipo ↓
Random	0.296	0.214	0.521
Ours	**0.294**	**0.209**	**0.515**

Table 9: Performance on validation sets of MoleculeNet regression tasks with different masking strategies.

Model	ESOL ↓	FreeSolv ↓	Lipo ↓
MolTRES gen	0.284	0.251	0.602
MolTRES-small disc	**0.280**	**0.250**	**0.594**

Table 10: Performance of the MolTRES generator and discriminator on validation sets of MoleculeNet regression tasks. Note that we use the discriminator of MolTRES-small, which has the same size of the generator of MolTRES.

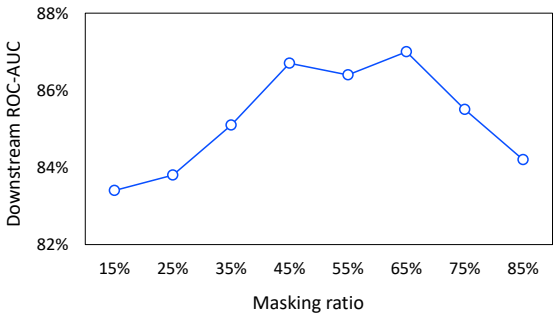

Figure 4: Comparison of MolTRES for different masking ratios on MoleculeNet classification tasks.

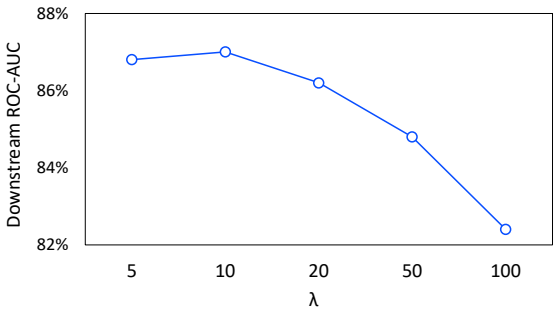

Figure 5: Comparison of MolTRES for different λ on MoleculeNet classification tasks.

A Appendix

Additional statistics and results. We report our analyses on MoleculeNet regression tasks in Tables 8 – 10. We also report dataset statistics in Tables 11 and 12.

Pre-training hyper-parameter analysis. We study the effect of pre-training hyper-parameters as shown in Figures 4 and 5. We report ROC-AUC scores on four MoleculeNet classification tasks (BBBP, ClinTox, BACE, and SIDER). First, in Figure 4, we find that the optimal masking ratio for MolTRES is 65%. When the masking ratio is smaller than 65%, we observe that the generator easily fills masked tokens, resulting in significantly biased labels towards original. In contrast, when the masking ratio is larger than 65%, we observe that there is few evidence in input SMILES tokens to predict their original molecules, leading to less effective training. In addition, in Figure 5, we identify that the optimal value of λ is 10, different from the original work on generator-discriminator training in NLP (Clark et al., 2020) using 50. We suspect that this is because SMILES modeling typically shows smaller losses from the generator than language modeling, and thus we need smaller λ to balance the generator and discriminator training.

Architecture analysis. We analyze diverse variations on the MolTRES architectures, particularly about the architecture of the generator and discriminator. We report ROC-AUC scores on eight MoleculeNet classification tasks and MAE scores on three MoleculeNet regression tasks from each variation. In Table 13, the architecture of our standard setting used in Section 4 is shown in (D). The variations in (A) denote training smaller MolTRES models, showing that reducing layers and hidden size show comparable performance degradation when their numbers of parameters are commensurate. Note that we choose to reduce layers, since it achieves faster model execution speed. The variations in (B) and (C) are about the architecture of generators. (B) contains the variations changing the hidden sizes while using the number of layers of the discriminator, while (C) contains the variations changing the numbers of layers while using the hidden size of the discriminator. In this comparison, we first observe that there is an optimal size of generators that generate training examples suitably challenging for discriminators. After empirical investigation, we choose to set the number of layers in the generator to half of that in the discriminator.

	Descriptions	# tasks	# samples
BBBP	Blood brain barrier penetration dataset	1	2,039
Tox21	Toxicity measurements on 12 different targets	12	7,831
ToxCast	Toxicology data for a large library of compounds	617	8,577
Clintox	Clinical trial toxicity of drugs	2	1,478
MUV	Maximum unbiased validation group from PubChem BioAssay	17	93,087
HIV	Ability of small molecules to inhibit HIV replication	1	41,127
BACE	Binding results for a set of inhibitors for $\beta-$ secretase 1	1	1,513
SIDER	Drug side effect on different organ clases	27	1,427

Table 11: Classification tasks from MoleculeNet.

	Descriptions	# tasks	# samples
QM9	12 quantum mechanical calculations of organic molecules	12	133,885
ESOL	Water solubility dataset	1	1,128
FreeSolv	Hydration free energy of small molecules in water	1	642
Lipophilicity	Octanol/water distribution coefficient of molecules	1	4,200

Table 12: Regression benchmarks from MoleculeNet.

	Generator		Discriminator		ROC-AUC ↑	MAE ↓
	# layers	Hidden size	# layers	Hidden size	(CLS)	(REG)
(A)	3		6		81.9	0.375
		512		512	82.2	0.371
(B)	12	384			83.6	0.341
	12	512			84.7	**0.336**
(C)	4				83.3	0.343
	8				84.5	**0.336**
	12				84.0	0.337
(D)	6	768	12	768	**84.8**	**0.336**

Table 13: Variations on the MolTRES architectures. Unlisted values are identical to those of the standard setting of MolTRES in (D). Following the experimental settings described in Section 4.1, ROC-AUC scores are measured on eight MoleculeNet classification tasks and MAE scores are measured on three MoleculeNet regression tasks.