Computers and Artificial Intelligence；Vol.2, No.3; 2025
E-ISSN: 3006 -3728 P-ISSN：3006-371X
DOI: https://doi.org/10.70267/ cai.25v2n3.150157
Published by: Zeus Press

MV-MoCL: A Multi-View Molecular Contrastive Learning Framework by Integrating SMILES, 2D Graph, and 3D Geometry

Hang Zhao[1, *], Huang Xu[1]

School of Information, Guizhou University of Finance and Economics, Guiyang, Guizhou, 550025, China

**Corresponding author: Hang Zhao*, E - mail: zzzzzh@mail.gufe.edu.cn*

Abstract

Machine learning-based molecular property prediction (MPP) has garnered significant attention in computer-aided drug discovery, with its primary goal being the accurate estimation of molecular properties using structural data to accelerate drug development. In recent years, multi-view-based molecular property prediction learning has aimed to integrate information from different molecular views to learn high-quality molecular representations. Furthermore, the three-dimensional geometric information of molecules encompasses a richer set of spatial features, which plays a critical role in enhancing the accuracy of molecular property prediction. However, existing models often overlook 3D information in molecules. To address this, we propose a contrastive learning model named MV-MoCL, which incorporates encoders for multiple molecular dimensions: a SMILES Transformer for the SMILES sequence view, a Graph Isomorphism Network (GIN) for 2D molecular graphs, and SchNet for 3D geometric conformations. By aligning representations from these views using a contrastive loss function, our approach captures rich, multi-faceted molecular features during pre-training, thereby improving performance on downstream molecular property prediction tasks and effectively mitigating the issue of scarce labeled data. The proposed model was evaluated on several benchmark datasets from MoleculeNet, and experiments demonstrate that MV-MoCL matches or surpasses existing models across multiple benchmarks.

Keywords

molecular property prediction, multi-view learning, graph neural networks, deep learning, contrastive learning, bioinformatics

1. Introduction

In modern drug discovery, accurate molecular property prediction serves as a fundamental research task(Deng et al., 2023). Enhancing prediction precision directly accelerates the drug development process and thereby helps pharmaceutical companies achieve greater returns. With the continuous advancement of deep learning, a primary objective in molecular property prediction is to leverage these technologies—along with drug-related data to accurately predict key molecular properties, including physicochemical characteristics, biological activity, and toxicity profiles (Guo et al., 2023, Li et al., 2022). With the continuous advancement of modern drug development technologies, the drug discovery cycle has been progressively shortened, leading to the ongoing enrichment of molecular structure databases. Therefore, a key challenge currently lies in how to effectively leverage these databases containing vast molecular structures to learn meaningful molecular

representations and uncover latent chemical information, thereby enhancing the accuracy of molecular property prediction.

Existing molecular property prediction (MPP) methods are primarily categorized into two main classes: single-view and multi-view approaches. In single-view MPP, a specific molecular representation is typically selected as the model input, such as molecular fingerprints (Kearnes et al., 2016, Wen et al., 2022), SMILES sequences (Honda et al., 2019, Wang et al., 2019), or 2D molecular graphs (Hu et al., 2019, Rong et al., 2020), to perform predictions for downstream tasks, including activity prediction, solubility prediction, and drug-target prediction. Single-view methods inherently prevent models from accessing information present in other molecular dimensions, such as 3D geometric and 2D structural information(Guo et al., 2020, Wu et al., 2023). This limitation results in molecular representations that are incomplete and lack comprehensiveness, thereby constraining the model's overall learning capacity . Compared to single-view methods, multi-view MPP integrates multiple molecular views—such as 2D molecular graphs, 3D conformations, SMILES sequences, and functional groups offering advantages in terms of information completeness and model robustness(Zhang et al., 2024, Wang et al., 2024, Liu et al., 2021, Lin et al., 2024). However, existing multi-view models are limited to fusing or concatenating only two molecular views and often neglect the 3D information. Additionally, labeled data in the field of molecular property prediction is relatively scarce, and acquiring high-quality molecular property data is costly; nevertheless, model training typically requires substantial amounts of data, rendering supervised learning models susceptible to overfitting. Therefore, learning highly generalizable molecular representations from unlabeled data through pre-training techniques has emerged as a key focus of current research.

To address the aforementioned challenges, this paper proposes a multi-view contrastive learning framework named MV-MoCL that integrates molecular 3D information. For the first time, it employs three encoders—SMILES Transformer(Honda et al., 2019), GIN (Xu et al., 2018), and SchNet (Schütt et al., 2017)—for contrastive learning, enabling the acquisition of molecular representations from one-dimensional sequence, two-dimensional topological, and three-dimensional geometric perspectives, respectively. Existing work, such as MolCLR (Wang et al., 2022), constructs augmented molecular graphs for contrastive learning by leveraging graph augmentation strategies including atom masking, bond deletion, and subgraph removal. However, for molecules, it is essential to consider the activity cliff problem, wherein minor structural modifications can lead to substantial changes in molecular properties. Therefore, during the pre-training phase, we primarily extract feature representations from diverse molecular encoders to ensure that molecular properties remain invariant (Jiang et al., 2024). Specifically, we capture molecular information from SMILES sequences using the SMILES Transformer, learn topological connections and local structural details via the GIN graph neural network, and acquire stereochemical information through SchNet. Our approach facilitates the learning of more comprehensive and robust molecular representations compared to single-view or simplistic fusion methods, thereby providing a powerful tool for data-driven drug design and materials discovery.
The main contributions of this paper are summarized as follows: We design a multi-view contrastive learning model that integrates molecular 3D information, leveraging large volumes of unlabeled molecular data for pre-training to acquire high-quality molecular representations. We conduct comprehensive experiments on multiple publicly available benchmark datasets, validating the effectiveness and superiority of our proposed model architecture.

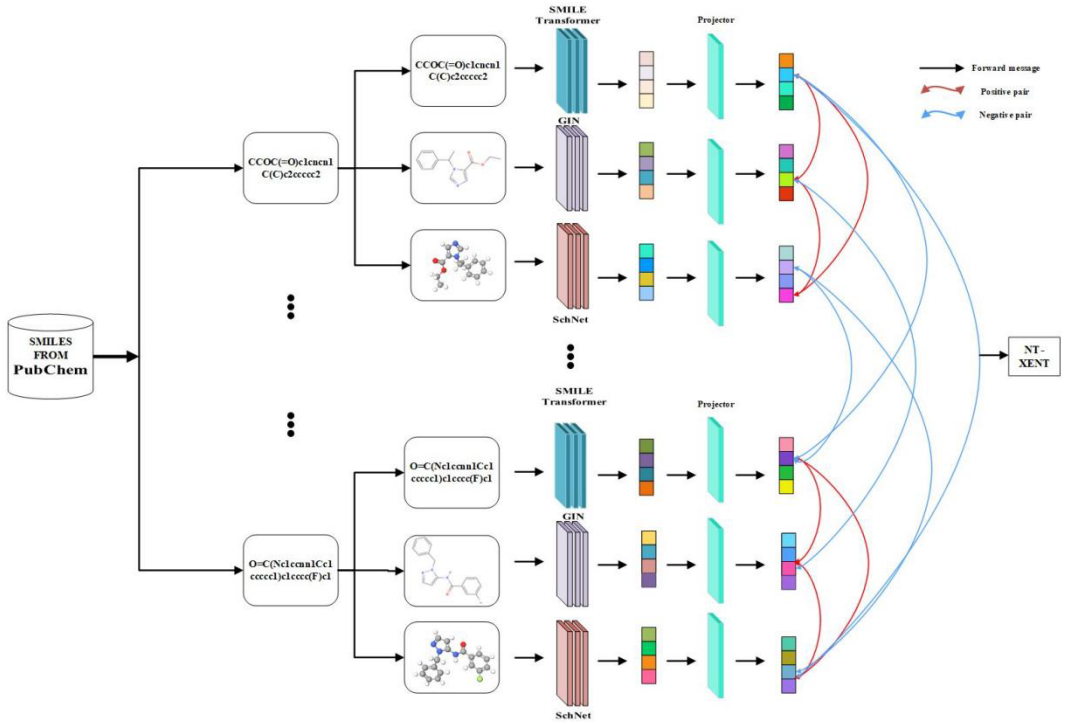
2. Methods

2.1 Overview of MV-MoCL

Figure 1 illustrates our model architecture. We select approximately 250,000 molecules with 3D geometric conformations from the PubChem database for pre-training. The same molecule is encoded separately by three encoders with distinct advantages—SMILES Transformer, GIN, and SchNet—followed by contrastive learning to aggregate the feature information of the same molecule. Specifically, we treat the features derived from the SMILES, 2D graph, and 3D geometric information of the same molecule as positive sample pairs, while features from different molecules serve as negative sample pairs. Through cross-view contrastive loss, the model learns that different views of the same molecule should be close in the feature space, whereas views from different molecules should be distant. Consequently, molecules with similar chemical properties will

form clusters in the feature space. Pre-training on large-scale unlabeled data enables the model to achieve superior generalization performance on downstream tasks.

Figure 1. Overview of MoleCFL.



2.2 Smile Transformer

The SMILES Transformer is a pre-trained model based on the Transformer architecture, specifically designed to extract molecular information from SMILES (Simplified Molecular Input Line Entry System) sequences. Drawing inspiration from how pre-trained models in natural language processing (NLP) handle language, it treats molecular structures as text sequences and captures molecular semantic information through unsupervised learning. Leveraging the SMILES Transformer enables efficient learning of relevant molecular representations from SMILES sequences without relying on structural representations in molecular graphs. The SMILES Transformer consists of multiple Transformer blocks, each comprising two core components: the self-attention mechanism and the feedforward neural network. The self-attention mechanism allows each element in the network to access contextual information from other elements in the sequence, while the feedforward neural network further processes this information to extract deeper-level features. In computing attention scores, we employ scaled dot-product attention. The computation of scaled dot-product attention can be represented by the following equations (1) and (2).

$$Attention(Q,K,V) = Softmax\left(\frac{Q\cdot K^T}{\sqrt{d_k}}\right)V \tag{1}$$

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}} \tag{2}$$

Where $\mathrm{Q} = \mathrm{XW^Q}, \mathrm{K} = \mathrm{XW^K}, \mathrm{V} = \mathrm{XW^V}$ are trainable parameters, $\mathrm{d_k}$ is the dimension of Q and K. To capture more information from the SMILES sequences, we employ multi-head attention, which enhances the model's understanding and processing capabilities for the data. Multi-head attention can be computed using equations (3) and (4).

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \dots, head_h)W^O \tag{3}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{4}$$

Here, h denotes the number of attention heads， W_i^Q, W_i^K, W_i^V, W^O is a new parameter matrix. After passing through n layers of Transformer blocks， we obtain the feature matrix for the SMILES string. Subsequently, we compute the mean and maximum values across all feature matrices and concatenate them to derive the molecular feature representation z^{1D}.

2.3 Graph Isomorphism Network

The Graph Isomorphism Network (GIN) is a model specifically designed for processing graph-structured data, aimed at addressing the limitations of graph neural networks in distinguishing different graph structures in terms of expressive power. It recursively updates node feature vectors by aggregating neighborhood node features, thereby enabling the differentiation of various graphs. In the field of molecular property prediction, GIN is widely employed to learn structural representations from 2D molecular graphs, effectively capturing the topological relationships between atoms and bonds within molecules. The strength of GIN lies in its ability to accurately aggregate and update node information by distinctly recognizing contributions from different neighbors, thereby preventing erroneous information propagation. Moreover, GIN incorporates an injective function during aggregation, which further enhances its capacity to learn powerful and discriminative representations of molecular structures. For a molecular graph $G = (V, E)$, where Vdenotes the set of atoms in the molecule and E represents the set of corresponding edges, the initial feature of each node is $h_v^{(0)}$. GIN integrates edge features into the node feature update equation (5), as shown below.

$$h_v^{(k)} = MLP^{(k)}\left((1+\epsilon^{(k)})\cdot h_v^{(k-1)} + \sum_{u\in\mathcal{V}(v)} ReLU(h_u^{(k-1)} + b_{uv})\right) \quad (5)$$

Where $h_v^{(k)}$ and $h_u^{(k-1)}$represent the embedding representations of atomic nodes v and u， b_{uv} denotes the embedding of the edge connecting nodes u and v， ϵ is a trainable parameter, and MLP is a multi-layer perceptron composed of multiple fully connected layers, employed to learn the aggregated neighborhood information, as shown in equation (6). Here,W_1, W_2, b_1, b_2 are learnable weights and biases.

$$MLP(x) = W_2 \cdot ReLU(W_1 \cdot x + b_1) + b_2 \quad (6)$$

$$z^{2D} = \sum_{k=0}^{K} READOUT\left(\{h_v^{(k)} \mid v \in V\}\right) \quad (7)$$

$$READOUT(H) = \frac{1}{|V|}\sum_{v\in V} h_v \quad (8)$$

Once all nodes in the molecule have been updated, we employ the average of all node representations as the output for the entire molecular graph z^{2D}, as shown in equations (7) and (8).

2.4 SchNet

Building upon prior research, we employ the SchNet model to learn spatial representations directly from the three-dimensional (3D) geometric information of molecules . Proposed by Schütt et al. in 2017, SchNet aims to model interatomic interactions within molecules for predicting total molecular energy, interatomic forces, and properties. This model is specifically designed to process molecular 3D geometries, enabling it to learn molecular representations directly from 3D atomic coordinates. It captures interatomic interactions in three-dimensional space through dynamically generated filters and consists of a series of hidden layers, as expressed in Equation (9) below:

$$h_i^{(k+1)} = MLP\left(\sum_{j=1}^{N} f_{CF}\left(h_j^{(t)}, r_i, r_j\right)\right) + h_i^{(t)} \quad (9)$$

Here, the input $h_i^{(0)}$ denotes the initial feature representation of atom v_i and f_{FG} represents the filter-generating network, as formulated in Equation (10).

$$f_{CF}(x_i, r_i, r_j) = x_j \cdot e_k(r_i - r_j) = x_j \cdot \exp\left(-\gamma \parallel r_i - r_j \parallel_2 - \mu \parallel_2^2\right) \quad (10)$$

$$z^{3D} = \frac{1}{N}\sum_{i\in v} h_i^{(K)} \quad (11) \quad (11)$$

Finally, the average of the node representations is adopted as the feature representing the 3D molecular geometry, as shown in Equation (11). Where K denotes the total number of hidden layers.

3. Experimental Setup and Results

3.1 Dataset

To evaluate the predictive capability of the MV-MoCL method, we conducted experiments on five benchmark datasets sourced from MoleculeNet (Wu et al., 2018), a benchmark for molecular machine learning established by the DeepChem team at Stanford University. The selected datasets include BBBP, BACE, ClinTox, Tox21, and SIDER .

Table 1. Description of the benchmark datasets.

Dataset	Tasks	Samples	Metric
BBBP	1	2039	ROC-AUC
BACE	1	1513	ROC-AUC
ClinTox	2	1478	ROC-AUC
SIDER	27	1427	ROC-AUC
Tox21	12	7831	ROC-AUC

BBBP: Binary labels of blood-brain barrier penetration(permeability).

BACE: Quantitative (IC50) and qualitative (binary label) binding results for a set of inhibitors of human β-secretase 1(BACE-1).

ClinTox: Qualitative data of drugs approved by the FDA and those that have failed clinical trials for toxicity reasons.

Tox21: Qualitative toxicity measurements on 12 biological targets, including nuclear receptors and stress response pathways.

Sider: Database of marketed drugs and adverse drug reactions (ADR), grouped into 27 system organ classes.

3.2 Baseline Models

To comprehensively evaluate the performance of our proposed MV-MoCL model, we compare it against seven other established methods in the field. A brief introduction of these baseline and state-of-the-art approaches is provided below:

ContextPred & AttrMask (Hu et al., 2019): This work introduced a self-supervised pre-training approach that constructs pre-training tasks at both the node level and the entire graph level. By simultaneously optimizing multiple pre-training objectives, it enables the model to learn richer representations .

GraphCL (You et al., 2020): This framework proposed four distinct types of graph data augmentation methods: node dropping, edge perturbation, attribute masking, and subgraph sampling. It further provided an experimental analysis of the effects of different data augmentation strategies .

JOAO (You et al., 2021): This study proposed a bi-level optimization framework that enables fully automatic and adaptive selection of data augmentation methods for graph contrast learning .

GraphMVP (Liu et al., 2021): This method proposed leveraging both the 2D topological structure and the 3D geometric view of molecules. It aims to learn richer molecular representations by maximizing the mutual information between these two views.

3D InfoMax (Stärk et al., 2022): This approach proposed maximizing the mutual information between the 2D and 3D vector representations of molecules, thereby equipping the model with the ability to extract potential 3D structural information from 2D molecular graphs .

G-Motif (Rong et al., 2020): This model proposed a fusion of the Transformer architecture with message-passing networks. It was pre-trained in a self-supervised manner on large-scale unlabeled molecular data .

3.3 Evaluation Metric

During the evaluation process, to maintain comparability with previous work, we adopted the area under the Receiver Operating Characteristic curve (ROC-AUC) as the primary metric. For each dataset, we conducted three independent runs using three distinct random seeds, recording the outcome of each run. The final experimental results are reported as the mean value along with its corresponding standard deviation calculated from these runs.

3.4 Implementation Details

During the pre-training phase, we employ a five-layer Graph Isomorphism Network (GIN) architecture and utilize the Adam optimizer for 100 training epochs with a batch size of 256. For the downstream fine-tuning stage, the Adam optimizer is retained for model optimization. All experiments are conducted on an RTX 2080 Ti GPU environment.

3.5 Results and Analysis

Table 2 summarizes the performance of the proposed MV-MoCL model on five benchmark datasets. The results demonstrate that our model achieves excellent performance across all downstream tasks, attaining state-of-the-art results on four of the five datasets. Calculation of the average performance reveals an improvement of 1.6%, underscoring the superiority of our proposed model. This also indicates that during the pre-training phase, our model successfully learns richer representations by integrating information from SMILES strings, molecular graphs, and 3D geometry. Compared to GraphMVP and 3D InfoMax, which utilize only the molecular graph and 3D geometric information, our model shows a significant average performance gain of 2.1% and 1.6%, respectively. Although our model's performance on the Tox21 dataset is slightly lower, this could be attributed to the inherent complexity of the Tox21 dataset and potential randomization factors.

Table 2. Results for molecular property prediction tasks. The best and second best results are marked bold and bold, respectively.

Dataset	BACE	BBBP	SIDER	Tox21	ClinTox	Avg
ContextPred	79.6(1.2)	64.3(2.8)	60.9(0.6)	75.7(0.7)	65.9(3.8)	69.2
AttrMasking	79.3(1.6)	64.3(2.8)	61.0(0.7)	**76.7(0.4)**	71.8(4.1)	70.6
GrapgCL	75.3(1.4)	69.7(0.7)	60.5(0.9)	73.9(0.7)	76.0(2.7)	71.0
JOAO	77.3(0.5)	70.2(1.0)	60.0(0.8)	75.0(0.3)	81.3(2.5)	72.7
GraphMVP	76.8(1.1)	68.5(0.2)	62.3(1.6)	74.5(0.4)	79.0(2.5)	72.2
3D InfoMax	79.4(1.9)	69.1(1.0)	60.6(0.7)	74.5(0.7)	79.9(3.4)	72.7
G-Motif	73.4(4.0)	66.4(3.4)	60.6(1.1)	73.2(0.8)	77.8(2.0)	70.2
MV-MoCL	**81.3(1.1)**	**70.3(1.0)**	**63.3(0.9)**	74.4(0.5)	**82.5(2.3)**	74.3

4. Conclusion

In this paper, we propose MV-MoCL, a multi-view contrastive learning-based pre-training framework that simultaneously integrates molecular information from multiple dimensions, including 1D SMILES strings, 2D molecular graphs, and 3D geometric structures. The key strength of our model lies in its ability to fully leverage the complementary nature of information across these different molecular views. Specifically, the SMILES Transformer captures sequential information from SMILES strings, the Graph Isomorphism Network (GIN) learns structural information from the 2D molecular graph, and SchNet extracts spatial conformation and stereochemical details from the 3D geometry. By integrating these diverse perspectives, MV-MoCL learns more comprehensive molecular representations. Experimental results demonstrate that our model outperforms existing methods on downstream molecular property prediction tasks, highlighting its strong generalization capability. Future work may explore incorporating additional molecular features—such as molecular fingerprints, functional groups, or knowledge graphs—during the downstream task phase to further enhance the model's generalization power and interpretability..

References

Deng, J., Yang, Z., Wang, H., Ojima, I., Samaras, D. and Wang, F., (2023). A systematic study of key elements underlying molecular property prediction. *Nature Communications,* vol. 14**,** no. 1, p. 6395.

Guo, Z., Guo, K., Nan, B., Tian, Y., Iyer, R. G., Ma, Y., Wiest, O., Zhang, X., Wang, W. and Zhang, C., (2023). Published. Graph-based molecular representation learning. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. IJCAI Press, pp. 6638-6646.

Guo, Z., Yu, W., Zhang, C., Jiang, M. and Chawla, N. V., (2020). Published. GraSeq: graph and sequence fusion learning for molecular property prediction. Proceedings of the 29th ACM international conference on information & knowledge management. ACM, pp. 435-443.

Honda, S., Shi, S. and Ueda, H. R., (2019). Smiles transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*.

Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V. and Leskovec, J., (2019). Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*.

Jiang, X., Tan, L. and Zou, Q., (2024). DGCL: dual-graph neural networks contrastive learning for molecular property prediction. *Briefings in Bioinformatics,* vol. 25**,** no. 6, p. bbae474.

Kearnes, S., McCloskey, K., Berndl, M., Pande, V. and Riley, P., (2016). Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design,* vol. 30**,** no. 8, pp. 595-608.

Li, Q., Zhang, Y. and Yan, J., (2025). ESG: Resource or Burden? Evidence from Chinese Listed Firms with Innovation Capability as the Mediating Mechanism. *Systems,* vol. 13**,** no. 9, p. 831.

Li, Z., Jiang, M., Wang, S. and Zhang, S., (2022). Deep learning methods for molecular representation and property prediction. *Drug Discovery Today,* vol. 27**,** no. 12, p. 103373.

Lin, J., Zheng, Y., Chen, X., Ren, Y., Pu, X. and He, J., (2024). Published. Cross-view Contrastive Unification Guides Generative Pretraining for Molecular Property Prediction. Proceedings of the 32nd ACM International Conference on Multimedia. ACM, pp. 2108-2116.

Liu, S., Wang, H., Liu, W., Lasenby, J., Guo, H. and Tang, J., (2021). Pre-training molecular graph representation with 3d geometry. *arXiv preprint arXiv:2110.07728*.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W. and Huang, J., (2020). Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems,* vol. 33, pp. 12559-12571.

Schütt, K., Kindermans, P.-J., Sauceda, H., Chmiela, S., Tkatchenko, A. and Müller, K.-R., (2017). Published. SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates, Inc., pp. 992-1002.

Stärk, H., Beaini, D., Corso, G., Tossou, P., Dallago, C., Günnemann, S. and Liò, P., (2022). Published. 3d infomax improves gnns for molecular property prediction. International Conference on Machine Learning. PMLR, pp. 20479-20502.

Wang, S., Guo, Y., Wang, Y., Sun, H. and Huang, J., (2019). Published. Smiles-bert: large scale unsupervised pre-training for molecular property prediction. Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics. ACM, pp. 429-436.

Wang, Y., Wang, J., Cao, Z. and Barati Farimani, A., (2022). Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence,* vol. 4**,** no. 3, pp. 279-287.

Wang, Z., Jiang, T., Wang, J. and Xuan, Q., (2024). Multi-modal representation learning for molecular property prediction: sequence, graph, geometry. *arXiv preprint arXiv:2401.03369*.

Wen, N., Liu, G., Zhang, J., Zhang, R., Fu, Y. and Han, X., (2022). A fingerprints based molecular property prediction method using the BERT model. *Journal of Cheminformatics,* vol. 14**,** no. 1, p. 71.

Wu, T., Tang, Y., Sun, Q. and Xiong, L., (2023). Molecular joint representation learning via multi-modal information of SMILES and graphs. *IEEE/ACM transactions on computational biology and bioinformatics,* vol. 20**,** no. 5, pp. 3044-3055.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K. and Pande, V., (2018). MoleculeNet: a benchmark for molecular machine learning. *Chemical science,* vol. 9**,** no. 2, pp. 513-530.

Xu, K., Hu, W., Leskovec, J. and Jegelka, S., (2018). How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

You, Y., Chen, T., Shen, Y. and Wang, Z., (2021). Published. Graph contrastive learning automated. International Conference on Machine Learning. PMLR, pp. 12121-12132.

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z. and Shen, Y., (2020). Published. Graph contrastive learning with augmentations. Advances in neural information processing systems. Curran Associates, Inc., pp. 5812-5823.

Zhang, R., Lin, Y., Wu, Y., Deng, L., Zhang, H., Liao, M. and Peng, Y., (2024). MvMRL: a multi-view molecular representation learning method for molecular property prediction. *Briefings in Bioinformatics,* vol. 25**,** no. 4, p. bbae298.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This work was supported by the GuiZhou University of Finance and Economics 2021 Undergraduate University-level Research Project [Grant Number 730321122901].

Copyrights