Enhancing Molecular Representation Learning Through the Combination of 3D and 2D Graph Machine Learning (Student Abstract)

Ian Tong Pan[1], **Joseph D. Romano**[1]

[1]University of Pennsylvania
{IanTong.Pan, Joseph.Romano}@pennmedicine.upenn.edu

Abstract

Molecular machine learning has broad applications across multiple domains such as drug development, environmental toxicology, and materials science. Various pre-trained frameworks using self-supervised representation learning have emerged to tackle the difficulty of obtaining large molecular datasets useful for training high-performing molecular machine learning models. In this study, we explore a novel representation learning framework trained using both 2D and 3D molecular data. Specifically, a 3D invariant graph neural network to learn how to capture 3D atomic information and then pass these atomic representations into a regular 2D graph neural network which can leverage molecular topology. Results from experiments demonstrate the representations produced by our method using both 3D and 2D molecular information lead to strong performance in downstream tasks.

Introduction

There have been numerous impactful applications of molecular machine learning, such as drug toxicity screening and chemical discovery (Hao, Romano, and Moore 2023; Guo et al. 2023). While a few graph neural network (GNN) based self-supervised representation learning models have been developed, there is still room for innovation for representation learning which is important as more informative representations may lead to better downstream task performance.

First, the vast majority of current models rely on two-dimensional (2D) chemical information (Guo et al. 2023). Incorporation of three-dimensional (3D) information can provide information about a molecule's biological activity and physicochemical properties. Additionally, positive pairs are defined unreliably in traditional contrastive learning frameworks. For example, a positive sample for a molecule would be created through augmentations such as bond deletion or subgraph removal, but these processes create new molecules that may have different characteristics (Wang et al. 2022).

Our paper introduces several contributions. First, we leverage both 3D and 2D information about a molecule through a novel dual-GNN architecture that leverages spatial arrangement as well as molecular topology information. Additionally, to avoid the concerns about previous contrastive learning approaches, we use different conformers, which are arrangements of atoms in a molecule generated from rotations of constituent groups about single bonds as positive samples, which maintain molecular structure.

Methods

The 3D invariant GNN used in our model is SchNet; originally designed for predicting molecular properties from atomic structures, it takes atomic charges and 3D atomic coordinates and applies continuous filter convolution layers, capturing the influence of neighboring atoms based on interatomic distances (Schütt et al. 2017). We chose SchNet because 3D invariance or equivariance to translations and rotations is required when dealing with spatial molecular systems. Additionally, it is more computationally efficient compared to other 3D invariant and equivariant GNNs. Moreover, we can safely pass the E(3) invariant node representations after 3 SchNet layers to our 2D GNN as the 2D GNN is not guaranteed to maintain equivariance.

After propagation through the 3D GNN, we append a global dummy node with an edge to each atom for each molecular graph which will hold the molecular representation. The atom node embeddings are taken from the 3D GNN. We used 4 GINEConv message passing layers to construct our 2D GNN. GINEConv extends the popular graph isomorphism operator to consider edge features, which is the bond type (e.g. single, double, triple, aromatic, and global) in our molecular graphs. In particular,

$$\mathbf{x}_i' = h_\Theta \left(\cdot \mathbf{x}_i + \sum_{j \in \mathcal{N}(i)} \text{ReLU}(\mathbf{x}_j + e_{j,i}) \right) \quad (1)$$

where x_i' is the updated node embedding of node i, x_i is the embedding vector of node i, x_j are the features of x_i's neighbors, $e_{j,i}$ is the edge feature between x_j and x_i, and h_Θ is a multi-layer perceptron in equation (1).

We use the temperature-scaled cross-entropy (NT-Xent) contrastive loss to train our model. Our pre-training dataset consists of the two lowest energy conformers from 309910 molecules retrieved from the QM9 and GEOM-Drugs datasets as positive pairs (Axelrod and Gomez-Bombarelli 2021; Chen et al. 2020). We use a stochastic gradient descent with momentum and weight decay optimizer and a cosine annealing with warm restarts learning rate scheduler.

Dataset	MACCS	ECFP	RDKit	HiMol	MolCLR	Our Method
BACE	0.719 (0.0408)	0.842 (0.0311)	0.847 (0.0317)	0.747 (0.0394)	0.622 (0.0441)	**0.828 (0.0325)**
BBBP	0.717 (0.0349)	0.644 (0.0357)	0.705 (0.0318)	0.525 (0.0382)	0.611 (0.0339)	**0.731 (0.0333)**
Clintox	**0.882 (0.0433)**	0.580 (0.0915)	0.570 (0.1038)	0.522 (0.0785)	0.669 (0.1030)	0.604 (0.0891)

Table 1: AUROC for MoleculeNet Classification Tasks (higher is better; standard deviation in brackets; best result in bold)

Dataset	MACCS	ECFP	RDKit	HiMol	MolCLR	Our Method
ESOL	1.741 (0.1552)	2.161 (0.1636)	1.981 (0.1498)	2.169 (0.1538)	1.770 (0.1332)	**1.597 (0.1123)**
Freesolv	**3.676 (0.4578)**	4.475 (0.5072)	4.332 (0.5282)	3.731 (0.4163)	4.180 (0.4892)	3.738 (0.4328)
Lipophilicity	1.425 (0.0432)	1.293 (0.0454)	1.344 (0.0404)	1.243 (0.0487)	1.187 (0.0440)	**1.151 (0.0410)**

Table 2: RMSE for MoleculeNet Regression Tasks (lower is better; standard deviation in brackets; best result in bold)

We compare our model against common cheminformatics fingerprints including MACCS, ECFP, and RDKit, as well as the embeddings generated by two other self-supervised learning approaches: MolCLR is a contrastive learning approach with only 2D molecular information, and HiMol, which uses a hierarchical pooling GNN based on functional groups, trained on a molecular reconstruction objective (Landrum 2006; Wang et al. 2022; Zang, Zhao, and Tang 2023). We train random forest classifiers and regressors (depending on the task) with molecular embeddings as input which are hyperparameter-tuned for each type of feature for a variety of benchmarks from the Moleculenet dataset, using their provided scaffold train, validation, and test splits(Wu et al. 2018). We evaluate task performance using area under the receiver operating characteristic curve (AUROC) for classification tasks and root mean squared error (RMSE) for regression tasks across 10000 bootstraps of the test set to obtain mean and standard deviation values.

Results

Our method demonstrates strong performance for multiple benchmarks as shown in Table 1 and 2. We believe this is due to our model's ability to incorporate 3D information into the molecular embeddings. For the binding affinity for human β-secretase (BACE) task, 3D information is useful because binding of molecules to proteins is dependent on the 3D conformation of the molecule. For blood-brain barrier penetration (BBBP), the ability of a molecule to cross the blood-brain barrier depends on its side, polarity, and hydrophobicity, which are all influenced by molecular spatial arrangement. In the same vein, water solubility (ESOL) and lipophilicity depend on molecular conformation, which provides logic for why our method is the best performer in these benchmarks. While solvation free energy (Freesolv) is dependent on molecular 3D structure, the dataset size is small compared to the other benchmarks with roughly 650 molecules total. This may explain why simpler embeddings perform better, as they are less likely to overfit. Clinical toxicity often depends on the presence of molecular functional groups, which may explain why incorporating spatial information did not improve performance on this dataset.

Conclusion

In this work, we have successfully introduced an algorithm for generating molecular embeddings that encode 3D information using GNNs. Our representation learning method can be applied for a variety of downstream cheminformatics tasks and may provide improved performance in tasks dependent on the spatial arrangement of molecules. Future work planned includes conducting further experiments with additional benchmarks and fine-tuning the model architecture to enhance performance.

References

Axelrod, S.; and Gomez-Bombarelli, R. 2021. GEOM.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *PMLR*.

Guo, Z.; Guo, K.; Nan, B.; Tian, Y.; Iyer, R. G.; Ma, Y.; Wiest, O.; Zhang, X.; Wang, W.; Zhang, C.; and Chawla, N. V. 2023. Graph-based Molecular Representation Learning. In *IJCAI-23*.

Hao, Y.; Romano, J. D.; and Moore, J. H. 2023. Knowledge graph aids comprehensive explanation of drug and chemical toxicity. In *CPT: Pharmacometrics Systems Pharmacology*.

Landrum, G. A. 2006. RDKit: Open-source cheminformatics.

Schütt, K. T.; Kindermans, P.-J.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; and Müller, K.-R. 2017. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *NeurIPS*.

Wang, Y.; Wang, J.; Cao, Z.; and Farimani, A. B. 2022. Molecular contrastive learning of representations via graph neural networks. In *Nature Machine Intelligence*.

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. In *Chemical Science*.

Zang, X.; Zhao, X.; and Tang, B. 2023. Hierarchical Molecular Graph Self-Supervised Learning for property prediction. In *Communications Chemistry*.