Integrating Expert Knowledge with Deep Learning Improves QSAR Models for CADD Modeling

Yunchao (Lance) Liu[1]，*Rocco Moretti*[2,3], *Yu Wang*[1]，*Bobby Bodenheimer*[1,4], *Tyler Derr*[1,5], *Jens Meiler*[2,3,6,7]

1. Department of Computer Science, Vanderbilt University, Nashville, TN 37212, USA

2. Department of Chemistry, Vanderbilt University, Nashville, TN 37212, USA

3. Center for Structural Biology, Vanderbilt University, Nashville, TN 37212, USA

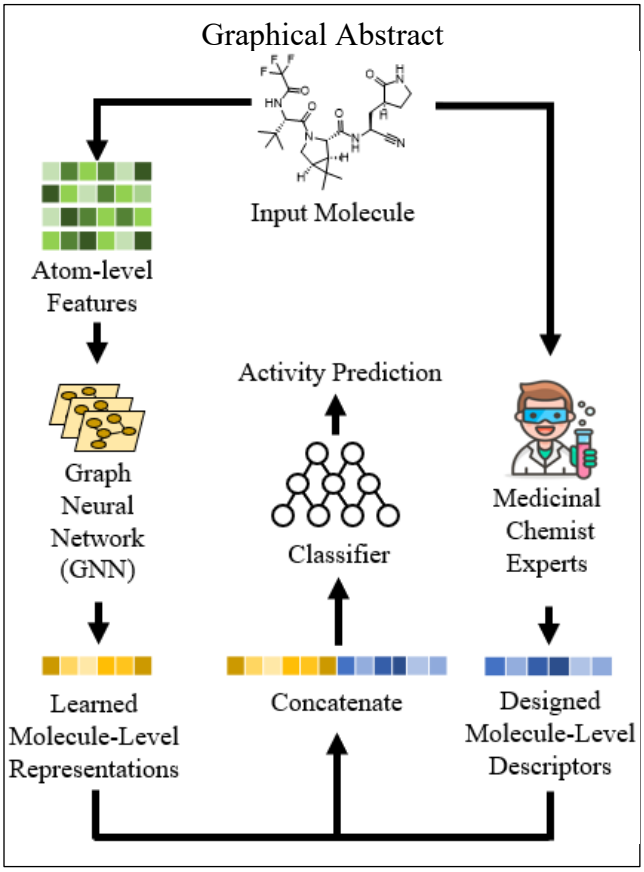
4. Department of Electrical Engineering and Computer Engineering, Vanderbilt University, Nashville, TN 37212, USA

5. Data Science Institute, Vanderbilt University, Nashville, TN 37212, USA

6. Institute of Drug Discovery, Leipzig University Medical School, 04103 Leipzig, Germany

7. Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), 04105 Leipzig, Germany

In recent years several applications of graph neural networks (GNNs) to molecular tasks have emerged. Whether GNNs outperform the traditional descriptor-based methods in the quantitative structure activity relationship (QSAR) modeling in early computer-aided drug discovery (CADD) remains an open question. This paper introduces a simple yet effective strategy to boost the predictive power of QSAR deep learning models. The strategy proposes to train GNNs together with traditional descriptors, combining the strengths of both methods. The enhanced model consistently outperforms vanilla descriptors or GNN methods on nine well-curated high throughput screening datasets over diverse therapeutic targets.



Introduction

Quantitative structure activity relationship (QSAR) modeling is one fundamental task in computer-aided drug discovery (CADD) [1]. In a machine-learning setting, traditionally expert-crafted descriptors are used to train a classifier to distinguish active and inactive compounds [2]. Numerous studies recently applied graph neural networks (GNNs) to molecule-related tasks, given the intrinsic graph nature of molecules [3-9]. GNNs can learn molecular representations

from molecular input data. Compared with the expert-crafted descriptors, the learned representations are believed to have several advantages over the fixed descriptors, including parsimony, interpretability, and task relevance [10, 11]. Even though previous studies have some promising results in GNN performance in molecular tasks, many of them either do not compare them with traditional approaches (e.g. [5, 12, 13] only compares their method with other GNN methods), or the task is not on QSAR modeling (such as quantum mechanics dataset QM9 used in [9]). Recently, some authors expressed doubts about deep learning performance over traditional methods in molecular tasks [14, 15]. Ultimately, it remains a mystery whether GNNs are consistently better than methods that rely on traditional descriptor in CADD [16, 17]. We hypothesize that these conflicting reports might invite an integrated method that combined GNNs with traditional descriptors to outperform both individual approaches, at least at the moment. Here we report such an integrated strategy. Evaluation on nine well-curated high throughput screening datasets demonstrates the effectiveness of this strategy, even over diverse therapeutic targets [18].

Method:

We train a neural network to predict activity combinig a GNN-derived molecular representation with the traditional expert-crafted descriptors [2]. Specifically, for a representation h from the GNN, we concatenate it with the descriptor h_{dp}.

$$h = GNN(m)$$

$$\hat{p} = f([h||h_{dp}])$$

where m is the input molecular graph and, h_{dp} is a descriptor. $f(\cdot)$ is a classifier, usually a Multi-Layer-Perceptron (MLP). $\hat{p}$ is the predicted activity.

The model is trained by optimizing the binary cross entropy loss L:

$$L = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{p}) + (1 - y_i)\log(1 - \hat{p})$$

where n is the number of samples in a batch, and y_i is the experimentally determined active/inactive status of the *i*-th molecule.

Experiments

- ***Datasets***

We validate the effectiveness of the proposed strategy via nine well-curated high-throughput screening (HTS) datasets. To avoid issues with experimental artifacts and high false positive rates [19], for the validation of our strategy, we chose datasets carefully curated [18] from high throughput screens in the PubChem database [20]. Only datasets with robust secondary validation of compounds were considered.

SMILES from the datasets were converted to SDF files using Open Babel [21], version 2.4.1. Standardized 3D coordinates are generated using Corina [22], version 4.3. Molecules are further filtered with atom type validity and duplicates with the BioChemical Library (BCL) [23].

Random split is used for the experiments, and each dataset is split into 80% for training and 20% for testing. Because preliminary results and previous literature [2] have shown that dropout can help avoid overfitting and the number of known active compounds is limited, we take the model from the last training epoch instead of the one from early stopping determined by validation performance. Multiple splits are used to prove the robustness of the proposed strategy.

- ***Evaluation Metric***

Logarithmic Receiver-Operating-Characteristic Area Under the Curve with the False Positive Rate in the range [0.001, 0.1] ($\text{logAUC}_{[0.001,0.1]}$): Ranged logAUC [24] is used because only a small percentage of molecules predicted with high activity can be selected for experimental tests

in consideration of cost in a real-world drug discovery campaign [19]. This high decision cutoff corresponds to the left side of the Receiver-Operating-Characteristic (ROC) curve, i.e., those False Positive Rates (FPRs) with small values. Also, because the threshold cannot be predetermined, the area under the curve is used to consolidate all possible thresholds within a certain small FPR range. Finally, the logarithm is used to bias towards smaller FPRs. Following prior work [2], we choose to use $\text{logAUC}_{[0.001,0.1]}$. A perfect classifier achieves a $\text{logAUC}_{[0.001,0.1]}$ of 1, while a random classifier reaches a $\text{logAUC}_{[0.001,0.1]}$ of around 0.0215, as shown below:

$$\frac{\int_{0.001}^{0.1} x \,\mathrm{d}\log_{10} x}{\int_{0.001}^{0.1} 1 \,\mathrm{d}\log_{10} x} = \frac{\int_{-3}^{-1} 10^{u} \mathrm{d}u}{\int_{-3}^{-1} 1 \mathrm{d}\, u} \approx 0.0215$$

Table 1. Dataset Statistics

Protein Target Class	**PubChem AID**	**Protein Target**	**Total Molecules**	**Active Molecules**
GPCR	435008	Orexin1 Receptor	218,156	233
	1798	M1 Muscarinic Receptor Agonists	61,832	187
	435034	M1 Muscarinic Receptor Antagonists	61,755	362
Ion Channel	1843	Potassium Ion Channel Kir2.1	301,490	172
	2258	KCNQ2 Potassium Channel	302,402	213
	463087	Cav3 T-type Calcium Channels	100,874	703
Transporter	488997	Choline Transporter	302,303	252

Kinase	2689	Serine/Threonine Kinase 33	319,789	172
Enzyme	485290	Tyrosyl-DNA Phosphodiesterase	341,304	281

- ***Results***

We used three GNN models in our experiments: GCN [25], SchNet [26], and SphereNet [9]. We used the BioChemical Library [27] to generate traditional QSAR descriptors. Following previous examples [2, 28], we use the optimal descriptors where 391-element molecular-level features are generated.

As can be seen from Table 2, BCL outperforms all three GNNs consistently. However, GNN performance is significantly augmented with the BCL descriptors in all nine datasets across all three GNNs. Moreover, although GCN, SchNet, and SphereNet perform differently, the enhanced models get similar results, demonstrating the robustness of this training strategy. In the following paragraph, we discuss the possible reasons for the effectiveness of this strategy.

Table 2. Results evaluated with $\text{logAUC}_{[0.001,0.1]}$. Values are average over five runs with different splits, with standard deviation.

PubChem AID	BCL Descriptor	GCN	SchNet	SphereNet	BCL +GCN	BCL +SchNet	BCL +SphereNet
435008	0.3±0.04	0.14±0.06	0.25±0.04	0.23±0.05	0.51±0.03	0.49±0.04	0.39±0.04
1798	0.22±0.04	0.12±0.04	0.2±0.04	0.21±0.04	0.3±0.06	0.33±0.05	0.31±0.03
435034	0.29±0.05	0.16±0.04	0.23±0.08	0.25±0.06	0.38±0.05	0.37±0.07	0.42±0.09

1843	0.45±0.04	0.23±0.05	0.39±0.06	0.34±0.06	0.65±0.08	0.59±0.08	0.59±0.08
2258	0.42±0.02	0.17±0.02	0.3±0.05	0.38±0.03	0.59±0.04	0.57±0.02	0.6±0.05
463087	0.35±0.03	0.14±0.01	0.29±0.03	0.33±0.03	0.49±0.03	0.58±0.03	0.62±0.03
488997	0.42±0.06	0.13±0.04	0.3±0.04	0.36±0.08	0.61±0.04	0.66±0.07	0.58±0.04
2689	0.53±0.05	0.34±0.06	0.4±0.04	0.51±0.08	0.63±0.05	0.58±0.07	0.61±0.06
485290	0.44±0.04	0.29±0.08	0.34±0.05	0.42±0.07	0.69±0.04	0.69±0.07	0.64±0.07
Average	0.38±0.01	0.19±0.01	0.3±0.01	0.34±0.01	0.54±0.01	0.54±0.02	0.53±0.01

The rationale for this approach is three-fold. First, data available for training in drug discovery campaigns is usually limited due to the high cost of experimental assays, and hence a way to increase usage efficiency is desired. The expert-crafted descriptors supplement GNNs with prior knowledge, i.e., descriptors that worked well in QSAR modeling in the past, which reduces the need for GNNs to learn that knowledge from a large amount of data. Secondly, GNNs typically have difficulty learning molecular-level features due to their limited receptive field or learning non-additive molecular-level features such as total polar surface area. On the other hand, molecular-level descriptors provide the global features directly. Thirdly, GNN intrinsically suffers from problems such as oversmoothing [29] and oversquashing [30] that introduce information loss in obtaining the global learned embedding from the atomic features. Meanwhile, the descriptors extract the molecular features directly and hence circumvent information loss, complementing GNN learned embeddings.

BCL features can be categorized into three types: 1). scalar 2). signed 2D autocorrelation [31] 3). signed 3D autocorrelation [31]. Thus, we did further analysis of each feature type's impact provided in the supplementary materials.

We note that a previous work ChemProp D-MPNN [32] used a similar strategy to boost their designed GNN in practice, further illustrating this method's practical value.

Future Work

The ablation study in the supplement investigates the contribution of different types of descriptors to the gained performance. Which of those features are practically unable to be learned by GNNs due to lack of data remains unanswered. We leave a detailed analysis of each descriptor feature to answer this for future work.

Conclusion

This letter introduces a simple but effective strategy to increase the GNN performance in QSAR modeling by combining the expert-crafted descriptors with GNN-learned representations via concatenation during training. We further validate the effectiveness of this strategy using nine well-curated HTS datasets. A comparison of before and after applying the strategy demonstrates the robustness of this strategy.

Corresponding Author

Tyler Derr: tyler.derr@vanderbilt.edu

Jens Meiler: jens.meiler@vanderbilt.edu

Author Contribution

Yunchao (Lance) Liu conceived and carried out the experiment and took the lead in writing the manuscript. Rocco Moretti contributed to the result interpretation and manuscript editing. Yu Wang contributed by providing an idea discussion. Bobby Bodenheimer contributed to

manuscript writing. Tyler Derr and Jens Meiler conceived the idea and contributed to the manuscript writing. Jens Meiler supervised the project and contributed to the final version of the manuscript. All authors provide critical feedback and help shape the research, analysis, and manuscript.

Funding Sources

Work in the Meiler laboratory is supported through NIH (R01 GM080403, R01 HL122010, R01 DA046138. J.M. is supported by a Humboldt Professorship of the Alexander von Humboldt Foundation. J.M. acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through SFB1423, project number 421152132 and through SPP 2363 for financial support.

Acknowledgment

Yunchao (Lance) Liu acknowledges that the Nvidia Academic Hardware Grant provides A6000 GPU for speeding up the computation.

References

1. Sliwoski, G., et al., *Computational methods in drug discovery.* Pharmacol Rev, 2014. **66**(1): p. 334-95.
2. Mendenhall, J. and J. Meiler, *Improving quantitative structure–activity relationship models using Artificial Neural Networks trained with dropout.* Journal of computer-aided molecular design, 2016. **30**(2): p. 177-189.
3. Klicpera, J., J. Groß, and S. Günnemann, *Directional message passing for molecular graphs.* arXiv preprint arXiv:2003.03123, 2020.
4. Zhang, Y., *An In-depth Summary of Recent Artificial Intelligence Applications in Drug Design.* arXiv preprint arXiv:2110.05478, 2021.
5. Wang, L., et al., *ComENet: Towards Complete and Efficient Message Passing for 3D Molecular Graphs.* arXiv preprint arXiv:2206.08515, 2022.
6. Liu, Y., et al., *Interpretable Chirality-Aware Graph Neural Network for Quantitative Structure Activity Relationship Modeling in Drug Discovery.* bioRxiv, 2022: p. 2022.08. 24.505155.
7. Schütt, K.T., et al., *Schnet–a deep learning architecture for molecules and materials.* The Journal of Chemical Physics, 2018. **148**(24): p. 241722.

8. Klicpera, J., et al., *Fast and uncertainty-aware directional message passing for non-equilibrium molecules.* arXiv preprint arXiv:2011.14115, 2020.
9. Liu, Y., et al., *Spherical message passing for 3d graph networks.* arXiv preprint arXiv:2102.05013, 2021.
10. Coley, C.W., et al., *Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction.* Journal of Chemical Information and Modeling, 2017. **57**(8): p. 1757-1772.
11. Duvenaud, D., et al., *Convolutional networks on graphs for learning molecular fingerprints.* arXiv preprint arXiv:1509.09292, 2015.
12. Rozemberczki, B., et al. *Chemicalx: A deep learning library for drug pair scoring*. in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2022.
13. Wang, Y., et al. *Imbalanced graph classification via graph-of-graph neural networks*. in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 2022.
14. Yu, Y., et al., *Do deep learning models really outperform traditional approaches in molecular docking?* arXiv preprint arXiv:2302.07134, 2023.
15. Zhou, G., et al., *Do Deep Learning Methods Really Perform Better in Molecular Conformation Generation?* arXiv preprint arXiv:2302.07061, 2023.
16. Jiang, D., et al., *Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models.* Journal of Cheminformatics, 2021. **13**(1): p. 12.
17. Kwapien, K., et al., *Implications of Additivity and Nonadditivity for Machine Learning and Deep Learning Models in Drug Design.* ACS Omega, 2022. **7**(30): p. 26573-26581.
18. Butkiewicz, M., et al., *Benchmarking ligand-based virtual High-Throughput Screening with the PubChem database.* Molecules, 2013. **18**(1): p. 735-56.
19. Butkiewicz, M., et al., *High-Throughput Screening Assay Datasets from the PubChem Database.* Chemical informatics (Wilmington, Del.), 2017. **3**(1): p. 1.
20. Kim, S., et al., *PubChem 2019 update: improved access to chemical data.* Nucleic acids research, 2019. **47**(D1): p. D1102-D1109.
21. O'Boyle, N.M., et al., *Open Babel: An open chemical toolbox.* Journal of cheminformatics, 2011. **3**(1): p. 1-14.
22. Gasteiger, J., C. Rudolph, and J. Sadowski, *Automatic generation of 3D-atomic coordinates for organic molecules.* Tetrahedron Computer Methodology, 1990. **3**(6): p. 537-547.
23. Benjamin P. Brown, O.V., Alexander R. Geanes, Sandeepkumar Kothiwale, Mariusz 4 Butkiewicz, Edward W. Lowe, Jr., Ralf Mueller, Richard Pape, Jeffrey Mendenhall, and Jens Meiler, *Introduction to the BioChemical Library (BCL): An application-based open-source toolkit for integrated cheminformatics and machine learning in computer-aided drug discovery.* 2022.
24. Mysinger, M.M. and B.K. Shoichet, *Rapid Context-Dependent Ligand Desolvation in Molecular Docking.* Journal of Chemical Information and Modeling, 2010. **50**(9): p. 1561-1573.
25. Kipf, T.N. and M. Welling *Semi-Supervised Classification with Graph Convolutional Networks*. arXiv e-prints, 2016. arXiv:1609.02907.
26. Schütt, K.T., et al., *Schnet: A continuous-filter convolutional neural network for modeling quantum interactions.* arXiv preprint arXiv:1706.08566, 2017.
27. Brown, B., et al., *Introduction to the BioChemical Library (BCL): An application-based open-source toolkit for integrated cheminformatics and machine learning in computer-aided drug discovery.* Frontiers in pharmacology, 2022: p. 341.
28. Vu, O., et al., *BCL::Mol2D-a robust atom environment descriptor for QSAR modeling and lead optimization.* J Comput Aided Mol Des, 2019. **33**(5): p. 477-486.

29. Chen, D., et al. *Measuring and relieving the over-smoothing problem for graph neural networks from the topological view*. in *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020.
30. Alon, U. and E. Yahav, *On the bottleneck of graph neural networks and its practical implications.* arXiv preprint arXiv:2006.05205, 2020.
31. Sliwoski, G., J. Mendenhall, and J. Meiler, *Autocorrelation descriptor improvements for QSAR: 2DA_Sign and 3DA_Sign.* Journal of computer-aided molecular design, 2016. **30**(3): p. 209-217.
32. Yang, K., et al., *Analyzing Learned Molecular Representations for Property Prediction.* Journal of Chemical Information and Modeling, 2019. **59**(8): p. 3370-3388.