Deep Neural Network Approach to Predict Properties of Drugs and Drug-Like Molecules

Magdalena Wiercioch[*]
Jagiellonian University
Department of Applied Computer Science
Faculty of Physics, Astronomy and Applied Computer Science
Łojasiewicza 11, 30-348 Kraków, Poland

University of Vienna
Department of Pharmaceutical Chemistry
Faculty of Life Sciences
1090 Vienna, Austria
mgkwiercioch@gmail.com

Johannes Kirchmair
University of Vienna
Department of Pharmaceutical Chemistry
Faculty of Life Sciences
1090 Vienna, Austria
johannes.kirchmair@univie.ac.at

Abstract

The discovery of small molecules with desirable properties is an essential issue in chemistry which could speed up much research progress in various domains such as virtual screening and drug design. Indeed, there is a series of open challenges, including building proper representations of molecules for machine learning algorithms. To address this issue, in this study we propose a deep neural network-based architecture that learns molecular representation to enhance the process of molecular properties prediction. We use two separate blocks of operations, where each block learns a representation. Then the two latent feature vectors are combined and fed into a few dense layers ended by a regression or classification layer. The performance of the proposed methodology was tested on the MoleculeNet, a standard benchmark for molecular machine learning. The results show that our method outperforms state-of-the-art models.

1 Introduction

Predicting molecular properties has attracted much attention in computer science, physics, or chemistry since it affects the speed of progress in discovering substances with desired characteristics for computer-aided drug discovery and materials development [1, 2]. Hence, computational methods, especially machine learning methods are more and more extensively used in many fields, including the exploration of vast numbers of synthetically accessible compounds [3].

Previous works have shown that the accurate modeling and prediction of molecule properties is strictly connected with the choice of molecular representation [4, 5]. Over the last few years, deep

[*] Corresponding author

Machine Learning for Molecules Workshop at NeurIPS 2020. https://ml4molecules.github.io

learning methods have made significant progress on numerous machine learning tasks. They have also revolutionized the way problems in cheminformatics are being solved [6]. Generally, the current works along the line of deep learning for molecules can be grouped into two categories: string- and graph-based methodologies. However, recent works indicate that graph representation and graph neural networks are a promising approach to tackle many challenges.

In this work, we introduce an architecture to predict molecular properties that comprises two separate blocks. Either block aims to learn an expressive representation from a given compound. The first block models a molecule as an undirected graph. The core part is a stack of attention layers followed by a fully connected layer applied to form features from molecular graphs. The second block converts a molecule into molecular fingerprints. Then a deep learning algorithm is adopted on the sequence to learn a representation. The two final feature vectors are concatenated and fed into fully connected layers. The final layer is a regression or classification layer to estimate the output as the property value.

A major advantage of our approach, as opposed to the previous methods lies in that our well-designed architecture applies a stacked attention mechanism and incorporates both the atom and molecule level attributes. Extensive experiments are provided on six datasets included in the publicly available benchmark dataset MoleculeNet [7]. Our approach significantly outperforms the state-of-the-art methods on both classification and regression tasks.

2 Model Architecture

Figure 1 shows the whole framework of our approach. The section which is located below the dotted line indicates the core part of the architecture. As can be seen from the figure, we employ two separate blocks to learn representations from molecules and combine these representations to feed into a series of fully-connected layers. At the ended we have a regression or a classification layer to estimate the output as the molecular property value or a prediction label (depending on the type of the task).

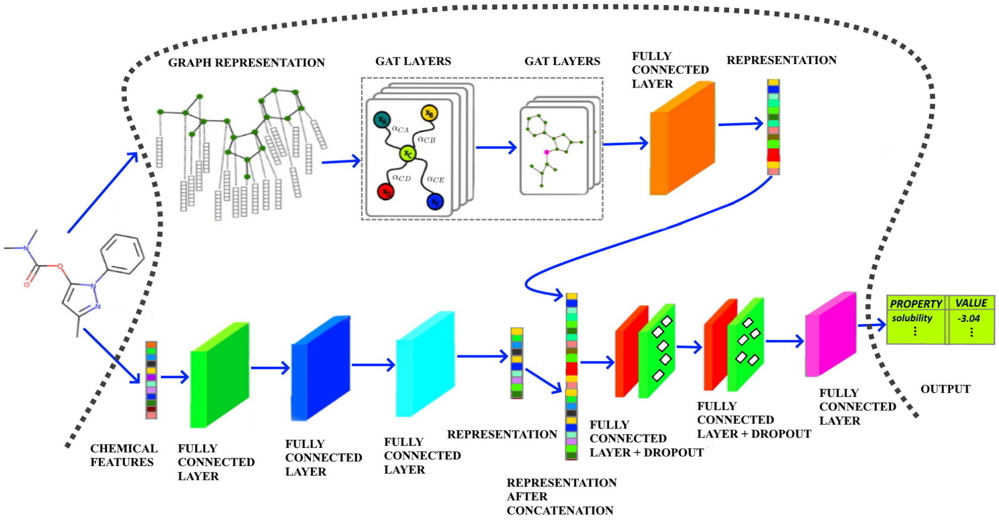
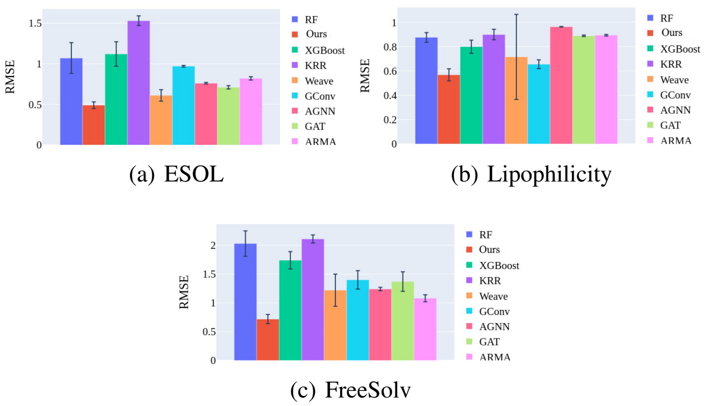


Figure 1: Depiction of the entire workflow of our approach.

Block I The first block begins by representing each chemical compound as an undirected graph containing nodes (atoms) with features $features_i$ and edges (bonds) with features $features_{ij}$. Formally, a molecule is denoted by $G = (\mathcal{V}, E)$, where $\mathcal{V}$ is a set of atoms containing $|\mathcal{V}| = N$ nodes. The graph is regarded as an undirected graph under the assumption that every atom has an interaction with others, including itself.

Then the model uses two separate GAT layers followed by a fully connected layer and a dropout layer. The GAT layers are marked with a dashed lines frame in Figure 1. The block returns a vector of features as output, x_{blockI}.

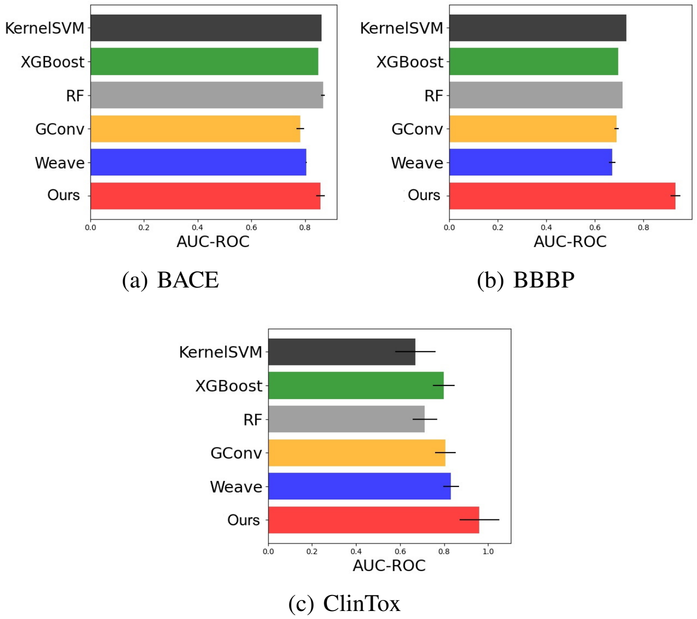


(a) ESOL

(b) Lipophilicity

(c) FreeSolv

Figure 2: The RMSE scores of various methods on regression task and test set. We achieved the least RMSE (lower is better).



(a) BACE

(b) BBBP

(c) ClinTox

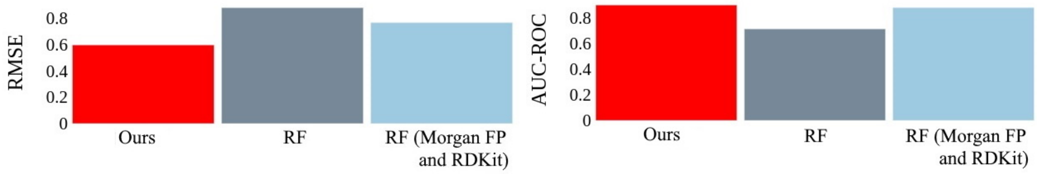
Figure 3: The AUC-ROC scores of various methods on classification task and test set. We achieved higher AUC-ROC score.

Block II Initially, the second block operates on a vector representation of features. Our goal is to obtain a neural fingerprint representation in order to ensure the generalization of molecular features [8]. Thus, we employ a collection of descriptors extracted by RDKit [9]. Then, the vector of features is fed into three fully connected layers and the block returns the representation, $x_{blockII}$.

The final molecular representation of a molecule x_{mol} is computed by concatenating the representations calculated at the blocks as follows: $x_{mol} = [x_{blockI} \cdot x_{blockII}]$.

3 Results and discussion

We evaluated the performance of our method on the ESOL, FreeSolv, Lipophilicity, ClinTox, BBBP, and BACE datasets from MoleculeNet [7]. Furthermore, each of the datasets was randomly split into training, validation and test sets at a ratio of 8:1:1. We performed 10 independent trials of training, validation and test, and averaged the outcomes. Furthermore, when we work on a regression task, we



(a) Lipophilicity (lower = better)

(b) BBBP (higher = better)

Figure 4: Performance of our approach and Random forest. Random forest baseline model is trained on the concatenation of MorganFPs and RDKit physchem descriptors as well.

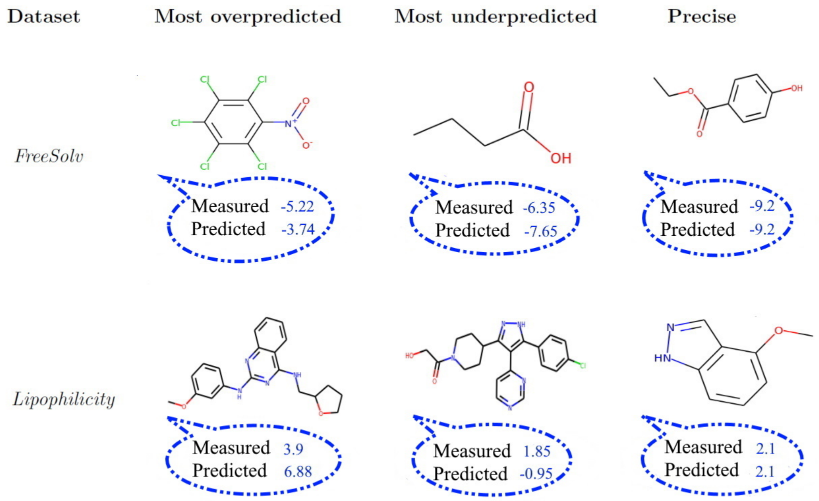


Figure 5: Exemplary molecules and corresponding most overpredicted, underpredicted and well-predicted values.

use root mean squared error (RMSE) as the performance metric. In turn, classification models are evaluated by the area under the receiver operating characteristic curve (AUC-ROC).

Figure 2 depicts the RMSE scores on the test sets. A glance at results reveals the proposed methodology as the model with the best performance for all datasets. Consequently, it achieves **0.49** over 0.61 (Weave), **0.71** over 1.22 (Weave), **0.56** over 0.65 (GConv) for ESOL, FreeSolv and Lipophilicity. Interestingly, our technique obtains the best AUC-ROC scores for the BBBP and ClinTox in the test dataset. For the BACE test dataset, the proposed method outperforms all benchmark models (0.858) except of KernelSVM (0.862). These results fully prove the validity of our approach. For comparison, we also evaluated the prediction performance where Random forest (RF) baseline model is trained on a feature vector generated by concatenating MorganFPs and RDKit descriptors. As illustrated in Figure 4, our architecture beats RF in all cases.

A small subset of molecules with the most underpredicted, overpredicted and well-predicted values is shown in Figure 5. One may notice that the maximum difference between the experimental value and the predicted one for a given molecule is not large. Moreover, most of the structures are different. It proves that our method can capture the molecular information and does not depend on the type of molecular structure.

4 Conclusions

We have presented a summary of the main results from a novel approach to predict the biological and physicochemical properties of small molecules. In the evaluation of our method with the MoleculeNet benchmark datasets, accurate prediction performance for various chemical properties were achieved.

Acknowledgments and Disclosure of Funding

This work is a part of broader research funded under the Iwanowska Program of the Polish National Agency for Academic Exchange (NAWA) - a decision no. PPN/IWA/2018/1/00094/DEC/1. This research was also supported by the National Science Centre (Poland) Grants No. 2016/21/N/ST6/01019. This publication has been partially funded from the SciMat Priority Research Area budget under the program "Excellence Initative - Research University" at the Jagiellonian University.

References

[1] E. N. Feinberg, R. Sheridan, E. Joshi, V. S. Pande, and A. C. Cheng, "Step change improvement in admet prediction with potentialnet deep featurization," *arXiv preprint arXiv:1903.11789*, 2019.

[2] K. Schütt, P.-J. Kindermans, H. E. S. Felix, S. Chmiela, A. Tkatchenko, and K.-R. Müller, "Schnet: A continuous-filter convolutional neural network for modeling quantum interactions," in *Advances in neural information processing systems*, pp. 991–1001, 2017.

[3] J. Levatić, M. Ceci, T. Stepišnik, S. Džeroski, and D. Kocev, "Semi-supervised regression trees with application to qsar modelling," *Expert Systems with Applications*, p. 113569, 2020.

[4] G. Cano, J. Garcia-Rodriguez, A. Garcia-Garcia, H. Perez-Sanchez, J. A. Benediktsson, A. Thapa, and A. Barr, "Automatic selection of molecular descriptors using random forest: Application to drug discovery," *Expert Systems with Applications*, vol. 72, pp. 151–159, 2017.

[5] K. V. Chuang, L. Gunsalus, and M. J. Keiser, "Learning molecular representations for medicinal chemistry," *Journal of Medicinal Chemistry*, 2020.

[6] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, "Concepts of artificial intelligence for computer-assisted drug discovery," *Chemical reviews*, vol. 119, no. 18, pp. 10520–10594, 2019.

[7] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "Moleculenet: a benchmark for molecular machine learning," *Chemical science*, vol. 9, no. 2, pp. 513–530, 2018.

[8] A. Gonczarek, J. M. Tomczak, S. Zaręba, J. Kaczmar, P. Dąbrowski, and M. J. Walczak, "Interaction prediction in structure-based virtual screening using deep learning," *Computers in biology and medicine*, vol. 100, pp. 253–258, 2018.

[9] G. Landrum *et al.*, "Rdkit: Open-source cheminformatics," 2006.