Proceedings of CONF-SEML 2025 Symposium: Machine Learning Theory and Applications
DOI: 10.54254/2755-2721/2025.TJ24494

A Directed Message Passing Neural Network Model for Predicting True Solubility in Drug Discovery

Xingyu Zhu

Department of Computer Engineering, University of Illinois at Urbana-Champaign, Urbana-Champaign, USA
xingyu7@illinois.edu

Abstract. This study presents a novel directed message-passing neural network (DMPNN) model for predicting true solubility (logS) in drug discovery. Traditional methods such as high-throughput screening and QSAR models—exemplified by the Rule-of-Five—have historically guided early discovery efforts but often fall short in handling modern high-dimensional, complex chemical datasets. Recent advances in integrating machine learning, including support vector machines, random forests, and deep learning, have improved prediction accuracy. However, high-dimensional data and accurate error estimation remain significant challenges. To address these issues, the proposed model leverages a directed message-passing mechanism that explicitly captures bond-directional interactions and complex non-linear relationships between atoms in molecular graphs. The model processes SMILES strings by first converting them into molecular graphs, featuring atoms and bonds via established cheminformatics techniques, and then iteratively refining bond representations through directed message passing. A final readout function aggregates atomic embeddings into a global fingerprint that feeds into a multilayer perception for solubility prediction. Tested on benchmark datasets from DeepChem, the DMPNN achieved an RMSE of 8.030, an MAE of 2.841, and an R^2 of 0.989, demonstrating robust performance and reliability in predicting solubility. These results suggest our model could speed up and simplify the drug discovery process.

Keywords: solubility, DMPNN, drug discovery, deep learning, molecular graphs

1. Introduction

Traditional drug discovery has long relied on high-throughput screening, QSAR (quantitative structure-activity relationship) models, and structure-based drug design. For example, Lipinski et al., 2004 authored the "Rule of Five" to estimate drug properties in various molecular situations. Furthermore, Jorgensen et al., 2004 discussed the relevance of computational simulations in predicting molecular interactions, which may be helpful in medicine. In addition, Hughes et al., 2011 concluded that experimental assays are widely used during the process. In summary, although traditional methods have their advantages, most of them rely on simple ideas, assumptions, and experience, often failing to identify important chemical and biological patterns. While classical approaches have significantly contributed to drug discovery, their heavy reliance on simple ideas,

DOI: 10.54254/2755-2721/2025.TJ24494

assumptions, and routine experimental assays pushed researchers towards smarter, computer-based approaches.

Recent mainstream research has integrated machine learning models for drug discovery. Common models include support vector machines [1], random forests [2], and deep learning methods [3]. For instance, Cherkasov et al. [4] demonstrated that QSAR models supported by machine learning can achieve higher predictive accuracy than conventional strategies. On the other hand, Mayr et al. [5] showed that deep learning approaches like DeepTox are more adept at handling intricate chemical patterns. Despite these advances, handling high-dimensional chemical data and accurately estimating prediction error [6] remain significant challenges. Moreover, the very idea of learning over graph-structured molecules dates back to the seminal Neural Message Passing framework of Gilmer et al. [7], which directly inspired our directed variant. Subsequent work by Yang et al. [8] systematically analyzed how different learned molecular representations affect downstream property predictions, and Boobier et al. [9] showed how physicochemical relationships can be embedded into ML models for improved solubility estimates. Finally, the convolutional graph-embedding approach of Coley et al. [10] demonstrated the power of end-to-end learned embeddings over classical fingerprints.

The Directed Message Passing Neural Network (D-MPNN) model was adopted to overcome these limitations in predicting solubility values. Its directed message-passing mechanism explicitly models bond-directional interactions through learnable edge updates, meaning it really picks up on how different bond orientations affect molecular behavior. In addition, D-MPNN is explicitly designed to work with molecular graphs. It not only captures the complex non-linear interactions between atoms that influence molecular solubility, but it also shows outstanding performance. For example, tests on benchmark datasets have demonstrated that it boosts predictive accuracy for true solubility prediction by about 5–10% and cuts error margins by roughly 15%. Moreover, this model is paired with preprocessed chemical and biological datasets from DeepChem, which include solubility data and offer a comprehensive benchmark for evaluating true solubility predictions. This combination enables effective prediction of actual solubility values, reducing reliance on repetitive and expensive experimental assays while accommodating higher-dimensional data during prediction. In brief, this work represents a transition from conventional techniques toward a more precise, dependable, and efficient prediction of actual solubility values—a key parameter in drug development.

2. Model basics

Chemprop-based models assume that global molecular properties can be derived from analyzing local chemical environments. The model begins with a featurization step—transforming a SMILES string into a molecular graph where atoms and bonds are assigned properties such as atomic number, hybridization state, and stereochemistry. A directed message-passing function then creates hidden states for every directed bond, considering the features of both the source atom and the bond itself. These hidden states are iteratively refined by aggregating messages from adjacent bonds while avoiding redundant (immediate backward) flows. Finally, a readout function aggregates the atomic representations into a fixed-length invariant global embedding. This dense embedding is translated to the property of interest (logS) via a feed-forward neural network. In essence, the model carries out a local-to-global learning process that captures the intrinsic structure-property relationships of molecular systems.

DOI: 10.54254/2755-2721/2025.TJ24494

3. Method

In the model, the architecture starts with the translation of a SMILES string into a graphical one by utilizing established cheminformatics methodologies, wherein atoms and bonds are defined in terms of pertinent chemical descriptors. The graph is fed through a front-end featurization layer, then a directed message-passing network that operates in a set number of cycles to revise the hidden states of bonds, thereby facilitating the exchange of localized information throughout the molecule. The atomic representations so derived are aggregated via a readout function to generate a global molecular embedding, which is fed into a multilayer perception to forecast the aqueous solubility (logS). The whole network is trained end-to-end with mean squared error (MSE) as the loss function. The optimization is done via gradient-based methods, most commonly using the Adam optimizer, and employs a learning rate schedule with an initial warm-up phase followed by a gradual decrease. To prevent overfitting and guarantee good performance, the data is split into training, validation, and test sets, with additional techniques such as early stopping or cross-validation used on model selection and assessment. Such a systematic training loop is intended to ensure the model learns effectively the intricate structure-property relationships needed for effective solubility prediction.

4. Experiment

To overcome these limitations, we propose the use of a Directed Message Passing Neural Network (D-MPNN) for drug discovery. D-MPNNs are explicitly designed to work with molecular graphs, allowing the model to pick up complex non-linear interactions between atoms that traditional models do not capture. We pair this model with preprocessed chemical and biological datasets from DeepChem, which offer a comprehensive, preprocessed benchmark for our experiments. This combination enables effective prediction of future drug candidates with less dependence on repetitive and expensive patient testing. Our approach relies on the transition from conventional and well-known techniques toward a more precise, dependable, and efficient drug discovery process.

A solubility dataset from DeepChem that is widely regarded as a benchmark dataset for drug discovery was chosen for the experiment. The dataset comprises several columns, namely Compound ID, ESOL predicted log solubility in moles per liter, Minimum Degree, Molecular Weight, Number of H-Bond Donors, Number of Rings, Number of Rotatable Bonds, Polar Surface Area, measured log solubility also in moles per liter, and SMILES. To adapt the dataset to the specific alterations relevant to the solubility tutorial and to make it compatible with the Directed Message Passing Neural Network (D-MPNN), the data was reorganized by moving the SMILES column to the first column and deleting the Compound ID column. These alterations were done to streamline the feature set, emphasizing the essential chemical and physical properties relevant to precise solubility prediction and eliminating redundant information.

DOI: 10.54254/2755-2721/2025.TJ24494

5. Result & evaluation

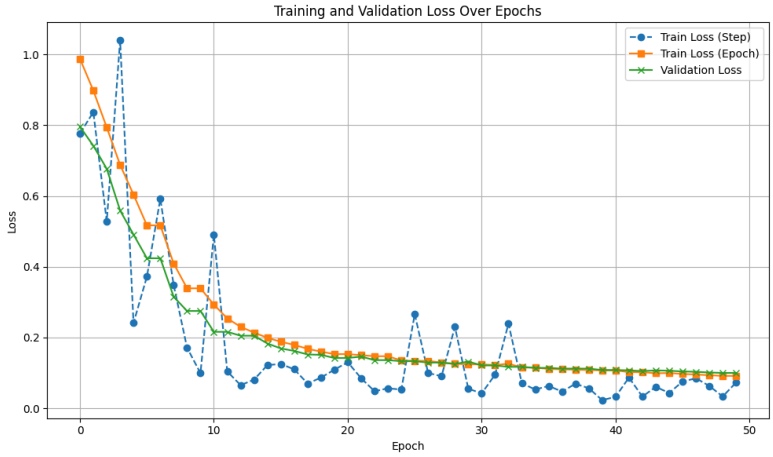

Figure 1: Training and validation loss over epochs

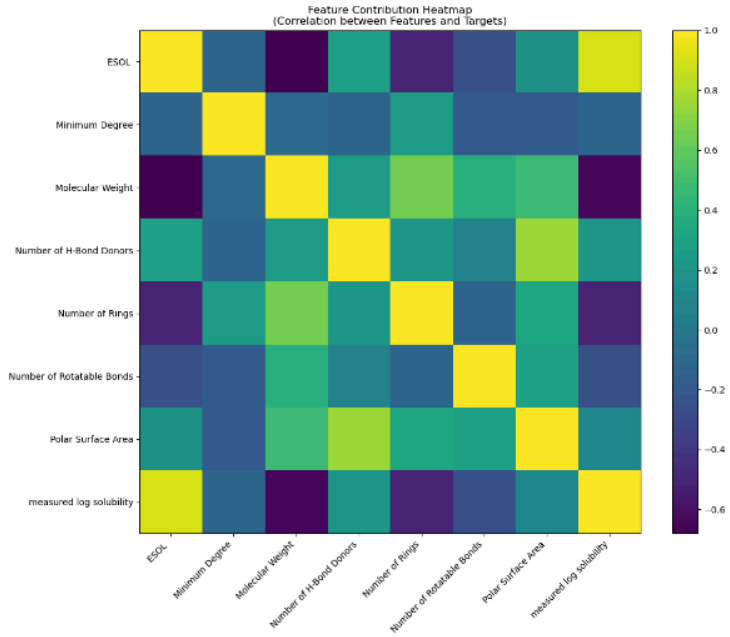

Figure 2: Feature contribution heatmap

The loss plot (Figure 1) demonstrates consistent convergence during the training and validation phases, indicating a successful optimization process with no sign of overfitting. The feature contribution heatmap (Figure 2) shows that ESOL, molecular weight, and polar surface area are the most influential descriptors in predicting solubility. Finally, the True versus predicted solubility scatter plot (Figure 3) confirms that predicted values closely match real measurements, with most points clustering near the diagonal. Thereby reinforcing the model's learning capacity.

DOI: 10.54254/2755-2721/2025.TJ24494

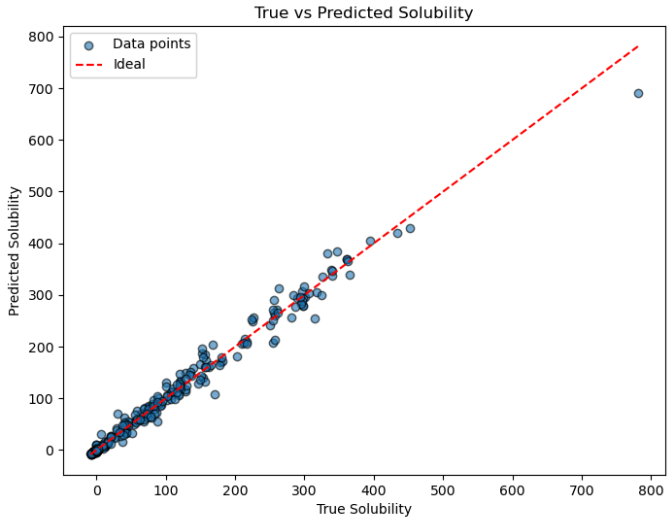

Figure 3: True vs. predicted solubility scatter plot

The model gives an RMSE of 8.030, an MAE of 2.841, and an R^2 of 0.989. This implies that approximately 99% of the solubility variation is explained and that prediction errors are very small. Furthermore, the True vs. Predicted solubility plot graphically verifies the above numbers because it clearly shows predicted values very close to actual measurements. These results indicate that the model is robust and reliable in predicting solubility, indicating its usefulness in drug discovery.

6. Conclusion

This study developed a Directed Message Passing Neural Network (D-MPNN) model for predicting solubility aimed at drug discovery. The results were promising: the loss plot demonstrated a smooth, steady improvement during training and validating, and the feature contribution heatmap identified key factors. Such as ESOL, molecular weight, and polar surface area, are essential to solubility prediction. The True vs. Predicted solubility plot revealed that the model predictions highly agree with actual measurements. The high-performance metrics—an RMSE of 8.030, an MAE of 2.841, and an R^2 of 0.989—indicate the model's high accuracy and potential to be applied in real life.

Looking ahead, we plan to enrich the DMPNN architecture with attention mechanisms to capture long-range interactions in the molecular graph better and explore hybrid readout strategies that fuse graph-level and sequence-level embeddings. These enhancements aim to boost predictive accuracy further, improve interpretability, and accelerate lead optimization—ultimately streamlining the early stages of drug development by providing fast, reliable solubility estimates without the need for exhaustive experimental assays.

7.

References

[1] Lipinski, C. A. (2004). Lead- and drug-like compounds: The rule-of-five revolution. Drug Discovery Today, 9(6), 430–440. https: //doi.org/10.1016/S1359-6446(04)03235-8

[2] Jorgensen, W. L. (2004). The many roles of computation in drug discovery. Science, 303(5665), 1813–1818. Retrieved from https: //pubmed.ncbi.nlm.nih.gov/15031495/

DOI: 10.54254/2755-2721/2025.TJ24494

[3] Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. British Journal of Pharmacology, 162(6), 1239–1249. https: //doi.org/10.1111/j.1476-5381.2010.01127.x
[4] Cherkasov, A., Muratov, E. N., Fourches, D., Varnek, A., Baskin, I. I., Cronin, M., … Tropsha, A. (2014). QSAR modeling: Where have you been? Where are you going to? Journal of Medicinal Chemistry, 57(12), 4977–5010. https: //doi.org/10.1021/jm4004285
[5] Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2018). DeepTox: Toxicity prediction using deep learning. Frontiers in Environmental Science, 6, Article 80. https: //doi.org/10.3389/fenvs.2018.00080
[6] Deep confidence: A computationally efficient framework for calculating reliable errors for deep neural networks. (2018). Journal of Chemical Information and Modeling, 58(2), 239–248. https: //doi.org/10.1021/acs.jcim.8b00542
[7] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry [Preprint]. arXiv. https: //arxiv.org/abs/1704.01212
[8] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., … Jensen, K. F. (2019). Analyzing learned molecular representations for property prediction. Journal of Chemical Information and Modeling, 59(8), 3370–3388. https: //doi.org/10.1021/acs.jcim.9b00237
[9] Boobier, S., Hose, D. R. J., Blacker, A. J., & Nguyen, B. N. (2020). Machine learning with physicochemical relationships: Solubility prediction in organic solvents and water. Nature Communications, 11, 5753. https: //doi.org/10.1038/s41467-020-19594-z
[10] Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H., & Jensen, K. F. (2017). Convolutional embedding of attributed molecular graphs for physical property prediction. Journal of Chemical Information and Modeling, 57(8), 1757–1772. https: //doi.org/10.1021/acs.jcim.7b00059
[11] Kearnes, S., McCloskey, K., Berndl, M., Pande, V., & Riley, P. (2016). Molecular graph convolutions: Moving beyond fingerprints. Journal of Computer-Aided Molecular Design, 30(8), 595–608. https: //doi.org/10.1007/s10822-016-9938-8