


Research Article

Predicting Drug Solubility Using Different Machine Learning Methods - Linear Regression Model with Extracted Chemical Features vs Graph Convolutional Neural Network

John Ho[†], Zhaoheng Yin[‡], Colin Zhang[§], Nicole Guo^{‡,§}, Yuwei Xia[#], Yang Ha^{§*}

Abstract

Predicting the solubility of given molecules remains crucial in the pharmaceutical industry. In this study, we revisited this extensively studied topic, leveraging the capabilities of contemporary computing resources by employing two machine learning models: a linear regression model and a graph convolutional neural network (GCNN) model. Using various experimental datasets, both methods yielded reasonable predictions. Despite its highest level of performance, the GCNN model has limited interpretability. On the other hand, although more human inputs and evaluations on the overall dataset is required, the linear regression model allows scientists for a greater in-depth analysis of the underlying factors through feature importance analysis. From the chemistry perspective, using the linear regression model elucidates the impact of individual atom species and functional groups on overall solubility, highlighting the significance of comprehending how chemical structure influences chemical properties in the drug development process. It has been learned that introducing oxygen atoms can increase the solubility of organic molecules, while almost all other hetero atoms except oxygen and nitrogen tend to decrease solubility.

Keywords: Drug design; Solubility; Linear regression model; Graph convolutional neural network; Feature importance.

Introduction

In the pharmaceutical industry, discovering new drugs is costly and time-intensive. An early-stage high-throughput screening (HTS) is usually used to reduce expenses and expedite the process by eliminating molecules that lack desired properties [1]. One key property is solubility, which governs drug uptake, movement, and metabolism in human bodies [2].

Prediction of molecular solubility, whether based on theoretical principles or experimental data, has been a prominent research field for decades. In 1968, Hansch et al. discovered that the octanol-water partition coefficient (P) can be used for solubility prediction [3]. Subsequently, the Yalkowsky group introduced a general solubility equation (GSE), which incorporated P and the melting point (MP) [4]. Later, Jorgensen and Duffy utilized Monte Carlo (MC) simulations to predict aqueous solubility by considering structural features such as molecular weight (MW), volume, solvent accessible surface area (SASA), hydrogen bond (HB) counts, and other physical descriptors like the solute–water Coulomb (ESXC) and Lennard–Jones (ESXL) interactions, as well as hydrophobic and hydrophilic components. Their approach achieved reasonable predictive accuracy on a dataset of 150 organic molecules [5].

Affiliation:

[†]Harvard University, Massachusetts Hall, Cambridge, MA, 02138

[‡]Berkeley Artificial Intelligence Research Lab, University of California Berkeley, Berkeley, California 94720

[‡]Carlmont High School, Belmont, CA, 94002

[§]Berkeley Center for Structural Biology, Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Laboratory, Berkeley, CA, 94720

[‡]Miramonte High School, Orinda, CA, 94563

[#]College of Education, The Pennsylvania State University, State College, PA, 16802

^{‡,†} Both the authors contributed equally.

*Corresponding author:

Yang Ha, Molecular Biophysics and Integrated Bioimaging, Lawrence Berkeley National Laboratory, Berkeley, CA, USA.

Citation: John Ho, Zhaoheng Yin, Colin Zhang, Nicole Guo, Yuwei Xia, Yang Ha. Predicting Drug Solubility Using Different Machine Learning Methods - Linear Regression Model with Extracted Chemical Features vs Graph Convolutional Neural Network. *Journal of Bioinformatics and Systems Biology*. 7 (2024): 92-97.

Received: January 23, 2024

Accepted: January 29, 2024

Published: March 06, 2024

In recent years, with the fast growth of computing power and the development of new algorithms, researchers can now work with more extensive datasets and employ sophisticated machine learning (ML) models [6]. Several databases, such as AQUASOL and PHYSPROP, used by Huuskonen et al. [7], ESOL by Delaney [8], and various solubility handbooks [9], have provided access to experimental solubility data for thousands of chemicals. AqSolDB is a newly developed database that combines multiple existing datasets [10]. From a methodological perspective, rather than relying on traditional regression models and classic neural network (NN) models, the Barzilay group applied graph convolutional neural networks (GCNN) for molecular property prediction. These GCNNs transform molecular structures into graphs, which can be input into a directed message-passing neural network, achieving state-of-the-art performance [11]. Moreover, research has extended beyond drug solubility in aqueous solutions to include solute types like small proteins [12] or various organic solvents [13].

While these advanced ML algorithms deliver remarkable performance, they often present challenges for human scientists seeking mechanistic insights into the chemistry behind these solubility models. These models are commonly believed to be "black boxes" because it remains difficult to understand the inner workings of, for instance, a 20-layer deep learning NN or a GCNN when all molecules are represented by extensive matrices. From a chemist's perspective, there is a growing need to shift the focus away from performance metrics and toward gaining deeper chemical insights. In this study, our goal is not to solely push the boundaries of predictive accuracy but to harness the strengths of both classical and

modern, sophisticated models to enhance our comprehension of the relationship between molecular structures and their chemical properties. With this knowledge, we aim to develop future ML models that combine high accuracy with human interpretability.

Methods

Two ML models were applied in this study: A linear regression model and a GCNN model (Figure 1).

In the linear regression model, we incorporated the molecular weight, total atom counts, and functional group counts as features to establish a multivariable regression with the experimental solubility values ($\log S$). The features were directly obtained from the molecular structure using the RDKit module [14] in SMARTS notation [15]. We also used L1 regularization with an alpha value of 0.01.

As for the GCNN, we employed the Chemprop model [16], which converts the atoms and bonds in the molecules into one-hot encoding, subsequently concatenating them into one tensor representing each individual atom or bond. Chemprop could construct three distinct tensors: one that maps each atom to its corresponding bonds (a2b), another that maps each bond to its corresponding atom (b2a), and a third that maps each bond to its reverse bond (b2revb). Subsequently, it combines each atom tensor into a unified vector and each bond tensor into another consolidated vector. Employing these five tensors, Chemprop identifies the neighboring bonds for each bond and aggregates their vector representations. Finally, the model appends this sum to the vector representations of both the bonds and atoms. These summated vector representations of individual bonds are then combined to generate one feature

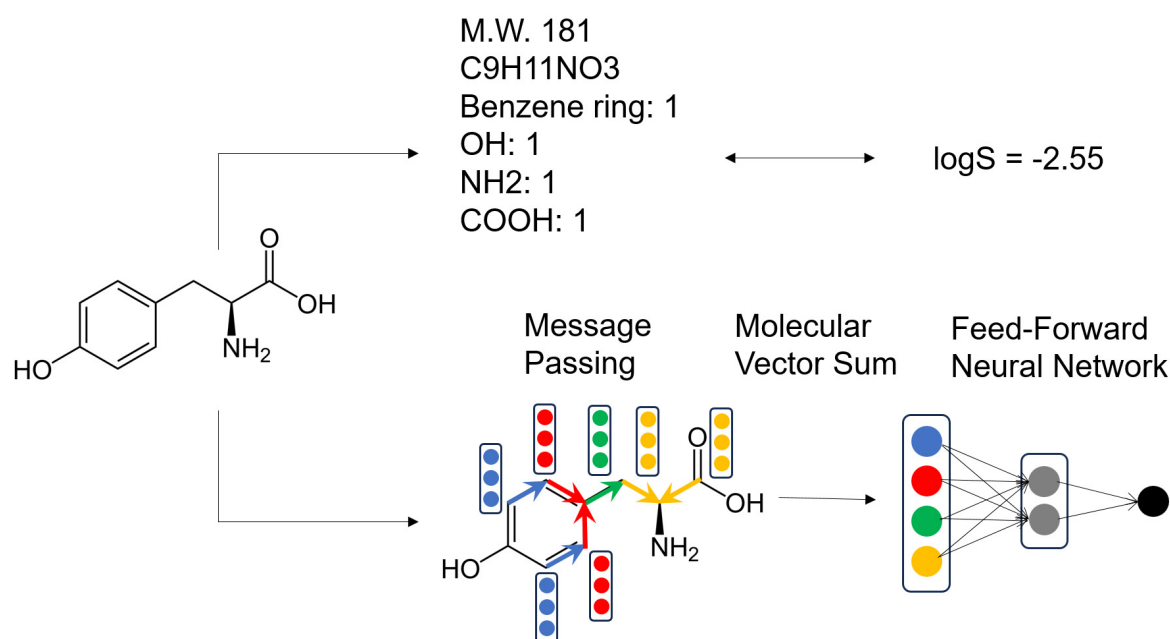


Figure 1: Configuration of the linear regression model (above) and GCNN model (below), using the tyrosine molecule as an example. The linear regression model relies on human-engineered features, including molecular weight (MW) and the count of functional groups, to predict experimental solubility ($\log S$), whereas the GCNN utilizes features acquired via message across a graph.

vector for the entire molecule, which enters a standard feed-forward neural network with a single output (logS).

The results from both models were tested on three different datasets: the Delaney, Huuskonen, and AqSolDB. The overall accuracy was evaluated against 5-fold cross-validation within each dataset, utilizing the root mean square error (RMSE) of the parity plots to assess the overall accuracy of the predictions.

Results

Predicting solubility

The parity plots for each model on different datasets are plotted in Figure 2, and the root-mean-square deviations (RMSE) are listed in Table 1.

Table 1: Performance of the Linear Regression Model and GCNN Model on Three Solubility Datasets

Dataset	Size	RMSE, Linear Regression Model	RMSE, GCNN Model
Delaney	1127	1.13	0.59
Huuskonen	1282	1.08	0.49
AqSolDB	9982	1.83	0.76

Across all three datasets, both the linear regression model and the GCNN model produced reasonably accurate predictions, with the majority of predicted values falling within 1 log unit of the actual values, consistent with findings in similar studies [17, 18]. Notably, both models exhibited their best performance on the Huuskonen dataset and the least optimal performance on the AqSolDB dataset.

By analyzing the outliers in Figure 2C, those data points correspond to ionic compounds such as compounds containing Zr^{4+} , Al^{3+} , and Zn^{2+} . This shows that the linear regression model has much poorer predictive power on those minorities that have quite different properties than the majorities. In this particular example, the majority of the dataset are neutral molecules. Imagine a human chemist will simply consider these ionic compounds to be very soluble while the regression model puts too much effort analyzing the organic part. The example suggests that the features for linear regression models need to be carefully determined and the composition of the datasets needs to be evaluated ahead. Here, a decision tree or a random forest model could be integrated to first filter out those ionic compounds that could greatly improve the performance of the linear regression model.

The GCNN model, on the whole, outperformed the linear regression model, particularly when dealing with larger and

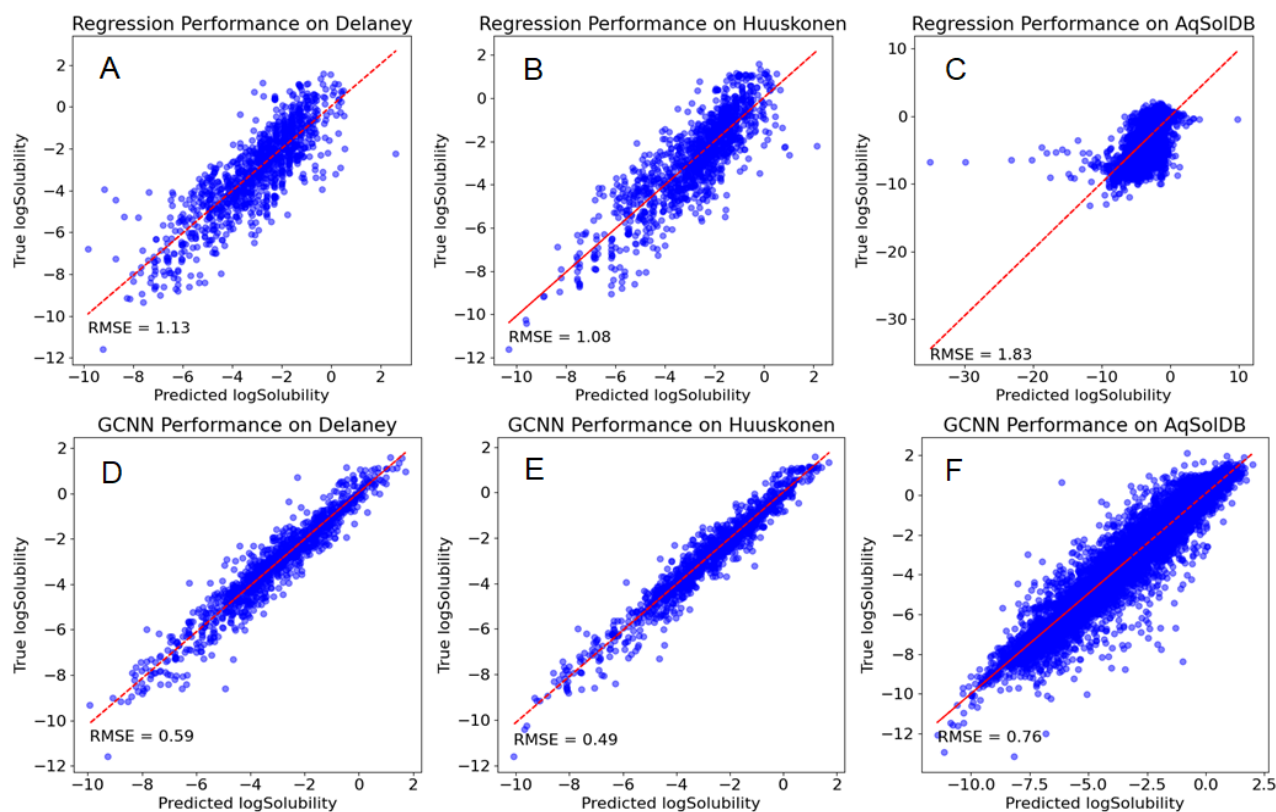


Figure 2: Parity plots for the Delaney, Huuskonen, and AqSolDB datasets using the linear regression model (above) and GCNN model (below). Predictions are shown from the validation folds of 5-fold cross-validation. Lines of best fit are shown in red.

diverse datasets, a result that aligns with the complexity and effectiveness of CNNs observed in various fields, including computer vision. However, this doesn't diminish the value of linear regression. Considering errors stemming from experimental conditions like pH and temperature, both models exhibit sufficient capabilities for drug design purposes.

Understanding the relationship between molecular structure and solubility

In contrast to the GCNN approach, which operates as a "black box", the linear regression model provides a relatively transparent depiction of the direct relationship between the input features and the solubility property of interest. Through feature importance analysis, we can visualize how each feature influences the final results. The significance of different atom species is presented in Figures 3.

Solubility hinges on the intermolecular forces between solute and solvent (water) molecules. In essence, polar molecules with more hydrogen bonds, whether as donors or acceptors, tend to exhibit higher solubility in aqueous solutions. The feature analysis results presented here provide a quantitative perspective on these conclusions. For instance, oxygen (O) atoms exert a strong positive influence on solubility because they not only increase the overall polarity of the organic molecules but also have the capacity to form hydrogen bonds with solvent water molecules. Conversely, halogens have a negative impact on solubility which can be quite counter intuitive (19). It is generally believed that halogen atoms, especially F and Cl can be hydrogen bond acceptors. However in reality, halogen atoms attaching to carbon chains could not form hydrogen bonds with water molecules. This underscores the pivotal role of hydrogen

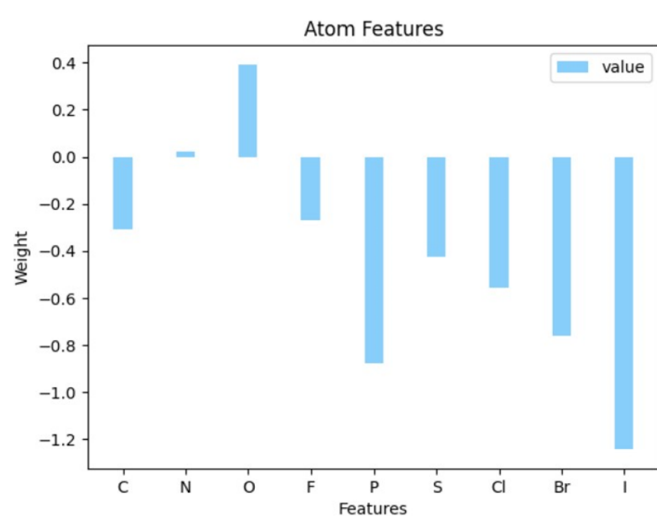


Figure 3: The linear regression weights of each type of atom feature for the Delaney dataset. Positive weights indicate features contributing to a relative increase in solubility, whereas negative weights indicate features that contribute to a relative decrease in solubility.

bonds in aqueous solubility, often surpassing the limited polarity they enhanced. It is also interesting to observe the trend of negative impact on solubility that $I > Br > Cl > F$, where heavier molecules are less likely to be soluble due to increasing London Dispersion Forces, while halogenated molecules are more likely soluble in hydrophobic solvents (20). This trait carries profound implications for drug delivery across cell membranes, making this extended exploration of solubility an area of considerable importance within the pharmaceutical field for further investigation.

Notably, the inclusion of functional group counts on top of atom counts yields a substantial improvement in the RMSE, as demonstrated in Table 2. This implies that the same type of atom when integrated into different functional groups, can exert varying effects on solubility. For instance, certain atoms like N, S, and P have the capacity to form diverse functional groups, which in turn may have either positive or negative impacts on solubility. The impact of different functional groups to aqueous solubility are shown in Figure S1.

It is essential to recognize that the machine learning models in this study can only predict a single solubility value for a given molecular structure. In reality, scientists contend with high-dimensional data encompassing a range of solubility values under varying conditions for each compound, as well as other physical and chemical properties. Tackling this complexity necessitates extensive data collection, cleaning, and algorithm development efforts. Ultimately, we anticipate that a sophisticated neural network-based model, coupled with interpretable feature analysis, will emerge as the preferred tool of choice, surpassing the simple linear regression approach.

From solubility prediction to drug design

As discussed above, simple solubility models have proven effective for high-throughput screening, even with the long-established GSE. Yet, the broader significance of solubility studies emerges in their capacity to inform and influence future drug design. This presents a reverse perspective: When endeavoring to create a drug molecule with specific solubility values or other desired physical attributes, the pivotal question becomes, which functional groups should be incorporated?

Table 2: Performance of the Linear Regression Model with Only Atom Feature and with Atom and Functional Group Features on Three Solubility Datasets

Dataset	Size	RMSE, Atom Features Only	RMSE, Atom, and Functional Group Features
Delaney	1127	1.13	0.96
Huuskonen	1282	1.08	0.89
AqSolDB	9982	1.83	1.73

Using the insights gained from feature importance analysis in this study, it is possible to develop a general understanding of which functional groups to incorporate. For instance, to enhance aqueous solubility, introducing an OH group to a side chain can be an effective strategy. At the same time, for improving the ability to permeate cell membranes, the inclusion of a halogen atom might be the most suitable choice. However, in real-world scenarios where multiple factors must be considered simultaneously, the complexity of human decision-making can be quickly overwhelmed. This is precisely where the GCNN model proves invaluable. By leveraging a well-trained neural network that establishes connections between defined molecular substructures and their associated properties, the coupling of the GCNN with a molecular generative model [21] has the potential to enable the generation of viable drug candidates with desired properties on a larger scale. This approach will likely drive the next generation of high-throughput screening in the pharmaceutical industry.

Conclusion

In this investigation, we tried to predict the drug molecules aqueous solubility by applying two distinct models: a linear regression model with human-engineered features, and a GCNN model. Both models exhibit commendable predictive accuracy across diverse datasets, with the GCNN delivering superior overall performance. Nonetheless, the linear regression model offers a valuable lens into the intricate interplay between specific features and solubility, shedding light on the significance of certain atoms, functional groups, and hydrogen bonds in the process. The integration of a GCNN model with feature analysis represents a promising avenue for future research in this domain.

Acknowledgments

We would like to thank Kyle Swanson for his valuable suggestions and feedback for this study. Yang Ha is supported by the Berkeley Center for Structural Biology (BCSB). Colin Zhang is supported by the Experiences in Research (EinR) program, which is funded by the Berkeley Lab Deputy Director for Research and Berkeley Lab Foundation.

References

- Pereira DA & Williams JA. Origin and evolution of high throughput screening. *Br J Pharmacol* 152 (2007): 53-61.
- Wen H, Jung H, Li X. Drug delivery approaches in addressing clinical pharmacology-related issues: opportunities and challenges. *AAPS J* 17 (2015): 1327-1340.
- Hansch C & Helmer F. Extrathermodynamic approach to the study of the adsorption of organic compounds by macromolecules. *J Polym Sci A Polym Chem* 6 (1968): 3295-3302.
- Yalkowsky SH & Valvani SC. Solubility and partitioning I: solubility of nonelectrolytes in water. *J Pharm Sci* 69 (1980): 912-922.
- Jorgensen WL & Duffy EM. Prediction of drug solubility from Monte Carlo simulations. *Bioorg Med Chem Lett* 10 (2000): 1155-1158.
- Llinas A & Avdeef A. Solubility challenge revisited after ten years, with multilab shake-flask data, using tight (SD~ 0.17 log) and loose (SD~ 0.62 log) test sets. *J Chem Inf Model* 59 (2019): 3036-3040.
- Huuskonen J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J Chem Inf Comput Sci* 40 (2000): 773-777.
- Delaney JS. ESOL: estimating aqueous solubility directly from molecular structure. *J Chem Inf Comput Sci* 44 (2004): 1000-1005.
- Yalkowsky SH, He Y, Jain P. *Handbook of Aqueous Solubility Data*. CRC Press (2016).
- Sorkun MC, Khetan A, Er S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Sci Data* 6 (2019): 143.
- Yang K, Swanson K, Jin W, et al. Analyzing learned molecular representations for property prediction. *J Chem Inf Model* 59 (2019): 3370-3388.
- Wirawan A, Harris RS, Liu Y, et al. HECTOR: a parallel multistage homopolymer spectrum based error corrector for 454 sequencing data. *BMC Bioinformatics* 15 (2014): 1-13.
- Chinta S, Rengaswamy R. Machine learning derived quantitative structure property relationship (QSPR) to predict drug solubility in binary solvent systems. *Ind amp Eng Chem Res* 58 (2019): 3082-3092.
- Landrum G. Rdkit: Open-source cheminformatics software (2016).
- Daylight Chemical Information Systems, Inc. SMARTS-A Language for Describing Molecular Patterns (2007). <https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html>
- Heid E, Greenman KP, Chung Y, et al. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J Chem Inf Model* 64 (2024): 9-17.
- Boobier S, Hose DR, Blacker AJ, et al. Machine learning with physicochemical relationships: Solubility prediction in organic solvents and water. *Nat Commun* 11 (2020): 5753.
- Ye Z, Ouyang D. Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms. *J Cheminform* 13 (2021): 1-13.

19. Huibers PD, Katritzky AR. Correlation of the aqueous solubility of hydrocarbons and halogenated hydrocarbons with molecular structure. *J Chem Inf Comput Sci* 38 (1998): 283-292.
20. AlSaleem SS, Zahid WM, AlNashef IM, et al. Solubility of halogenated hydrocarbons in hydrophobic ionic liquids: Experimental study and COSMO-RS prediction. *J Chem Eng Data* 60 (2015): 2926-2936.
21. Merz Jr KM, De Fabritiis G, Wei GW. Generative models for molecular design. *J Chem Inf Model* 60 (2020): 5635-5636.