Molecular property prediction using Graph Neural Networks
CIS*6190 - Machine Learning for Sequences

Gaëlle Quillaud
Department of Computer Science
University of Guelph / Télécom Saint-Étienne
Guelph, Canada / Saint-Étienne, France
gquillau@uoguelph.ca

Abstract

This paper presents an approach to predict aqueous solubility from molecular structure using Graph Neural Networks (GNNs). Traditional methods in drug discovery often rely on handcrafted features, which limit accuracy due to fixed input sizes. On the other hand, GNNs offer a data-driven approach, extracting features directly from raw inputs. Molecules are represented as graphs, where nodes represent atoms and edges represent bonds.
The solution involves preprocessing the data, selecting node and edge features, and training the Graph Convolutional Network model. The results demonstrated the effectiveness of GNNs in capturing features from the molecular structure and predicting the solubility.

1 Introduction

Molecular property prediction is a fundamental challenge for drug discovery, where the goal is to accurately predict molecular properties, such as bioactivity, solubility, and toxicity, based on their chemical structures. High precision in molecular property prediction reduces the experimental cost and time of drug development.

Using Machine Learning for predicting molecular properties significantly saves resources and time, by reducing the number of clinical trials (1). It enables the analysis of molecular data to identify potential drug targets, reducing the number of molecules that need to be tested and also suggesting new types of molecules that researchers might not have considered.

2 Background

2.1 Traditional methods

Traditional machine learning methods for molecular property prediction and more broadly for drug discovery often rely on handcrafted features because the models can only handle inputs of a fixed size (2), which prevents information from being learned directly from raw inputs and limits the accuracy (3).
For instance, the dataset used in this paper is derived from the research paper *ESOL: Estimating Aqueous Solubility Directly from Molecular Structure* (J.S.Delaney) (4), which uses hand-crafted features, including molecular weight, proportion of heavy atoms in aromatic systems, number of rotatable bonds, and others. The relationship between features and solubility is estimated using multiple linear regression. This involves fitting a linear equation to the training data, where the coefficients represent the influence of each feature on the predicted solubility.

In other domains, Deep Learning has had very successful results in the past years, but its application remains limited in the field of drug discovery and cheminformatics. Several commonly used neural network architectures have been tested in this field and improved the results obtained by traditional Machine Learning methods (1) such as Support Vector Machines and Random Forests. However, they still rely on handcrafted features and are not specifically adapted to the structure of data like molecules.

2.2 Graph Neural Networks in cheminformatics

2.2.1 Representing molecules as Graphs

Graphs are data structures representing objects and the relation between them. Those objects are

called vertices, or nodes, and the relations are called edges. A Graph is notated $G = (V, E)$ where V is a set of nodes and E is a set of edges.
A Graph is fully determined by its adjacency matrix A, which represents the connections between nodes, A_{ij} being equal to the number of connections from node *i* to node *j*. For a non-oriented Graph, i.e. a Graph where if there is a link from node *i* to node *j*, there is also one from node *j* to node *i*, the adjacency matrix is symmetric.

As molecules are a set of atoms with some of them linked to others, they can be represented by Graphs, where nodes represent atoms and edges represent bonds. Their adjacency matrix is symmetric because the bond between two atoms is not oriented.
An example of the Graph representation of a molecule is shown in Figure 1.

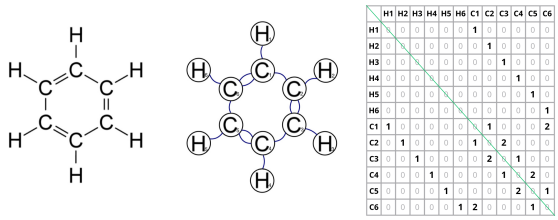


Figure 1: Graph and Adjacency matrix of the Benzene molecule

2.2.2 Graph Neural Networks

Neural Networks have shown very good results for grid-structured data like images. One of the most used architectures for handling images is Convolutional Neural Network (CNN). However, unlike images, graphs have irregular shapes and sizes; there is no spatial order imposed on the nodes, and the number of neighbors changes whereas in an image, a non-edge pixel has exactly 8 neighbors. Thus, generalizing CNNs from images to graph input is a challenge, as classical convolutional operations cannot be applied to graphs.

To face this issue, Gori, Monfardini, and Scarselli introduced Graph Neural Networks, which extend neural network architectures to handle Graph-structured data (5). Similarly to traditional Convolutional Neural Networks, Graph Convolutional Networks (GCNs) learn features by inspecting neighboring nodes. They aggregate node vectors, pass the result to the dense layer, and apply non-linearity using the activation function (6), but the method of convolution differs from that of a classical Convolutional Neural Network.

As molecules can be represented as graphs, Graph Neural Networks became one of the most promising Machine Learning models for cheminformatics. They have been introduced in this field by D. Duvenaud, D. Maclaurin, et al. in *Convolutional Networks on Graphs for Learning Molecular Fingerprints* (2). This work showed the effectiveness of Graph Neural Networks in adapting to the task at hand by using data-driven features instead of hand-crafted ones, and outperformed or matched the accuracy of circular fingerprints across various tasks.

Used in drug discovery, Graph Neural Networks enable automatic feature extraction from raw inputs and are capable of capturing the structural features of molecules in a more comprehensive manner compared to traditional methods.

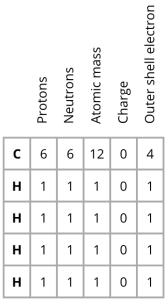
3 Methodology

Solubility is the ability of a substance, called solute, to dissolve into another substance, called solvent, to create a homogeneous solution. Aqueous solubility is one of the key molecular properties for drug discovery, as about 40% of drugs are insoluble in water and thus fail to reach the market (7). Because of this importance, accurately predicting solubility is fundamental but remains challenging. This paper aims to address this issue, using the dataset from the paper *Estimating Aqueous Solubility Directly from Molecular Structure* (4), where the solubility is estimated using linear regression. The dataset is composed of 2874 molecules and four columns for each one: its name, the measured solubility (mol/L), the ESOL predicted solubility, and the SMILES representation of the molecule.

3.1 Feature selection

There are two types of Graph features: node features and edge features. They characterize the atoms and the bonds of the molecule and impact its properties. They have to be chosen carefully depending on the property we intend to predict. For instance, node features can be the atom's symbol, its number of protons and neutrons, its

degree, its charge, its number of radical electrons, its hybridization, its aromaticity, its number of connected hydrogens, whether it is chiral or not, etc. Edge features can be the bond's type (simple, double, triple), the conjugation, the stereo, whether it is part of a ring or not, etc.



	Protons	Neutrons	Atomic mass	Charge	Outer shell electron
C	6	6	12	0	4
H	1	1	1	0	1
H	1	1	1	0	1
H	1	1	1	0	1
H	1	1	1	0	1

Figure 2: Example of node features for the methane molecule CH_4

The two main factors influencing solubility are the geometry and the polarity of the molecule (8). Indeed, molecules with larger surfaces tend to have higher solubility because they can interact with solvent molecules over a greater area, leading to stronger solute-solvent interactions. The overall shape of a molecule also affects its solubility by determining its ability to fit into the solvent environments. For instance, molecules with complementary shapes to the solvent molecules are more likely to dissolve readily.
Polarity is the second property that significantly impacts the molecule's solubility in polar solvents such as water. Polar molecules have regions of partial positive and partial negative charges. Water molecules surround the polar molecules, aligning their dipoles to maximize hydrogen bonding interactions. Nonpolar molecules, which do not have dipole moments, have minimal interactions with water molecules and tend to be insoluble or poorly soluble in water.
Therefore, several features from both node and edge properties can impact solubility, mainly by impacting the geometry or the polarity of the molecule:

1. Node Features:

- **Atom's Symbol**: The type of atom influences its ability to interact with solvent molecules. For example, polar atoms like oxygen or nitrogen tend to form hydrogen bonds with water, increasing solubility.
- **Degree**: The degree of an atom (number of bonds it forms) can affect solubility by influencing the molecule's overall polarity and surface area available for interactions with the solvent.
- **Charge**: Charged atoms or ions can interact with polar solvents via electrostatic interactions, affecting solubility. For instance, ions with opposite charges can dissolve in polar solvents due to attraction.
- **Hybridization** (9): Hybridization refers to the mixing of atomic orbitals to form hybrid orbitals that are used by atoms for bonding. Hybridization influences the geometry and polarity of the molecule, which can influence its interactions with solvent molecules.
- **Aromaticity** (11): Aromaticity refers to a property that certain cyclic molecules with a specific arrangement have. Aromatic compounds are characterized by exceptional stability. They often have unique solubility properties due to their delocalized electron clouds. They can interact differently with solvents compared to non-aromatic compounds, affecting solubility.
- **Number of Connected Hydrogens**: The presence of hydrogen atoms can affect the polarity of a molecule, influencing its solubility.

2. Edge features:

- **Bond's Type**: Different bond types (single, double, triple) can influence the overall polarity and structure of molecules, affecting their interactions with solvents and thus solubility.
- **Conjugation** (11): Conjugation refers to double bonds separated by one single bond. Conjugated systems can affect the distribution of electron density within a molecule, impacting its polarity and interactions with solvents, thereby influencing solubility.
- **Stereochemistry and Chirality** (11): Stereochemistry depicts the three-dimensional arrangement of atoms in molecules. Molecules are chiral if they cannot be superimposed on their mirror image. A molecule's 3D geometry can affect its shape and polarity, influencing its solubility properties.
- **Ring Membership**: Molecules containing ring structures can have different solubility behaviors than acyclic molecules due to differences in rigid-

ity, molecular shape, and interactions with solvents.

3.2 Data preprocessing

3.2.1 SMILES Representation of molecules

In most molecular databases, molecules are represented with Simplified Molecular-Input Line-Entry System (SMILES) strings.

The SMILES (12) representation is built upon the Graph representation of the molecule without hydrogen atoms. It is obtained by going through the Graph using a depth-first traversal algorithm and printing the symbol of each traversed node (atom) successively. First, the cycles are opened to transform the Graph into a spanning tree. When a cycle is opened, a numerical suffix is added to indicate the connection of the vertices corresponding to the removed chemical bond. Parentheses are used to indicate branching points on the tree.

Due to the nature of this language, there are multiple possible representations for the same molecule, depending on the starting atom and the order of traversal of the molecular structure. For instance, CCO, OCC, $C(C)O$, and $C(O)C$ are all valid representations of the ethanol molecule CH_3–CH_2–OH.

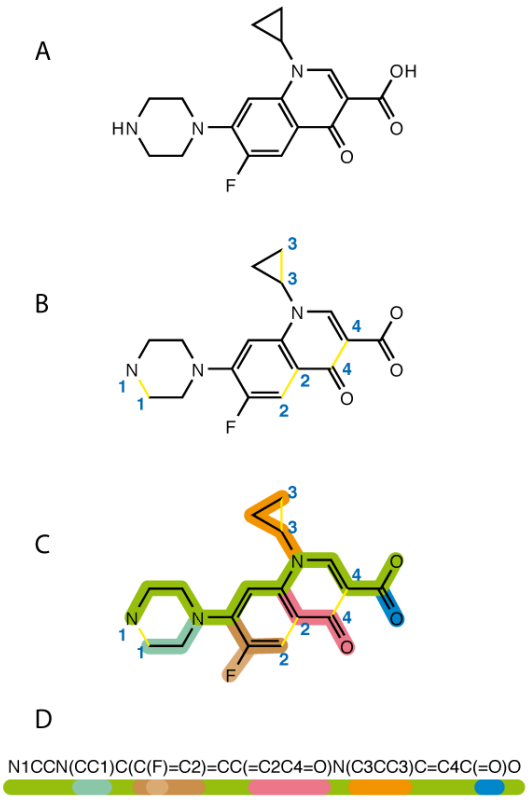


Figure 3: SMILES generation algorithm for the ciprofloxacin molecule. **A.** Structure of the molecule. **B.** Spanning tree obtained by removing hydrogen atoms and opening cycles. **C.** Tree covering. **D.** Corresponding SMILES representation.

3.2.2 Conversion to an RDKit object

While the SMILES representation of the molecule contains information about its structure, it doesn't provide information about the atoms' and the bonds' properties.

To get that information, the SMILES string was used to generate a molecule object, using the open-source toolkit for cheminformatics RDKit. This object contains information about the molecule, its atoms, and its bonds. In particular, the chosen features can be extracted from this object.

3.2.3 Conversion to a Graph

To use the molecule as the input of a Graph Neural Network, it has to be converted to a Graph that can be used in PyTorch Geometric, a library built upon PyTorch to manage Graph Neural Networks.

The conversion involves iterating over the bonds and adding them to the Graph as edges and iterating over the atoms and adding them to the Graph as nodes. Features are extracted from atoms and bonds at every step and added to the Graph.

The resulting structure is a Graph that contains the structure of the molecules, as long as edge and node features that can be used for the learning process of the Graph Neural Network.

3.3 Graph Neural Network

The model used is a Graph Convolutional Neural Network, implemented using PyTorch Geometric, a library built upon PyTorch to manage GNNs. The layer-wise propagation formula used by PyTorch Geometric is from the paper *Semi-supervised Classification with Graph Convolutional Networks* (13) and is the following:

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \qquad (1)$$

Where $\tilde{A} = A + I_N$ is the adjacency matrix of the connected graph with added self-connections, $\tilde{D}_{ii} = \Sigma_j \tilde{A}_{ij}$, $W^{(l)}$ is a layer-specific trainable weight matrix, $\sigma(\cdot)$ is the activation function,

such as ReLu, and $H^{(l)} \in \mathbb{R}^{N \times D}$ is the matrix of activations in the l^{th} layer; $H^{(0)} = X$.

The matrix $\tilde{D}$ applies an average operation to scale down the features, ensuring the feature calculation is robust against the effect of the number of connected nodes (otherwise, the feature value of a node that is connected to just one other will be lower than the one of a node connected to more than one other).
Multiplying the adjacency matrix $\tilde{A}$ with the node features computes the average features of the neighboring nodes for each node.
The weight matrix W creates a new representation of the vector of the current node.

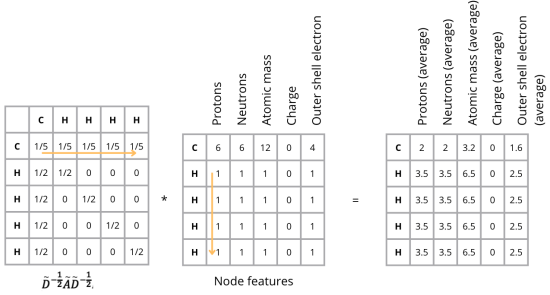


Figure 4: Formula applied to the input node features vectors X for the methane molecule

This computation is applied to all nodes of the graph and creates new node feature vectors for all nodes. Each time the node feature is processed, the corresponding node gets information about its adjacent nodes only. So if we want the node to get information about further nodes, we need to add more convolutional layers.

The GCN model consists of multiple convolutional layers, followed by a global max pooling that aggregates the node features into a graph-level representation. Experiments have been made on the number of convolutional layers and the size of those layers. The input layer has an input size of 43, which corresponds to the number of features, and the output layer has an input size of 1 since predicting solubility is a regression task.

4 Results

The success of the project has been measured based on the performance of the GNN model in accurately predicting molecular properties. Classical regression evaluation metrics such as mean squared error (MSE) and mean absolute error (MAE) have been used to assess this performance.
The model has been trained on a training set composed of 80% of the data set and evaluated on a test set composed of 20% of the data set.

4.1 Influence of the number and size of convolutional layers

Table 1 shows the influence of the number of convolutional layers on the average MSE loss on the test data, trained for 100 epochs. The loss represents an average of the losses obtained for hidden layer sizes of 8, 16, 32, and 64. The results presented in Table 1 illustrate the GCN model's sensitivity to variations in the number of layers.
Generally, increasing the number of layers tended to improve the predictive performance for a small

1 layer	2 layers	3 layers	4 layers	5 layers	6 layers	7 layers
3.812	3.470	3.297	3.191	2.914	6.012	5.950

Table 1: Influence of the number of convolutional layers on the MSE loss on test data

	2 layers	3 layers	4 layers	5 layers	6 layers	7 layers
8	3.575	3.288	3.032	2.849	2.726	2.643
16	3.459	3.252	3.108	2.813	2.721	2.658
32	3.392	3.244	3.383	2.843	3.152	2.853
64	3.453	3.405	3.241	3.151	15.45	15.65
128	4.376	4.454	3.118	67.08	45.37	1166850
256	3.693	4.167	18.24	148.3	9329	8E+08
512	4.319	4.745	9.791	110411	6E+06	5E+10

Table 2: Influence of the number of convolutional layers and of the size of those layers on the MSE loss on test data

convolutional layer size (less than 32), as shown in Table 2. However, for larger layer sizes, the loss was exponentially higher. These findings show that deeper GCN architectures may be more effective for predicting aqueous solubility directly from molecular structure, as long as the size of the hidden layers is small. The computational time increasing with the number of layers, the selected architecture was composed of 5 hidden convolutional layers of size 16. Adding a linear layer at the end of the convolutional neural network didn't show significant changes.

4.2 Final model and comparison with the original paper

After some research on the best Graph Convolutional Network architecture for predicting a continuous value such as solubility, the final chosen model was inspired by the Graph Convolutional Network from the paper *Attention-Based Graph Neural Network for Molecular Solubility Prediction* (14). The model is composed of two graph convolutional layers followed by two fully connected layers, with ReLU activation functions applied after each layer, and a final linear layer for the regression prediction. Dropout regularization is applied after the first graph convolutional layer to prevent overfitting and global max pooling is used to aggregate node features into a graph-level representation before going through the fully connected layers. Figure 5 shows the evolution of the loss during the training over the epochs.

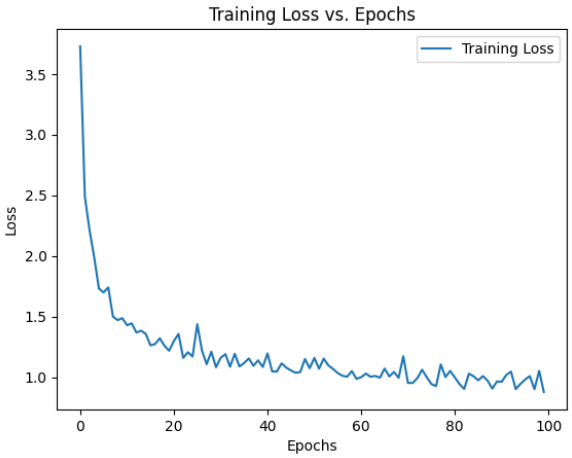


Figure 5: Training Loss vs. Epochs

The model demonstrated good results on the test set, with an average MSE loss on test data of 0.952 and an average MAE loss on test data of 0.751. Figure 6 shows the predicted solubility values versus the actual solubility values.

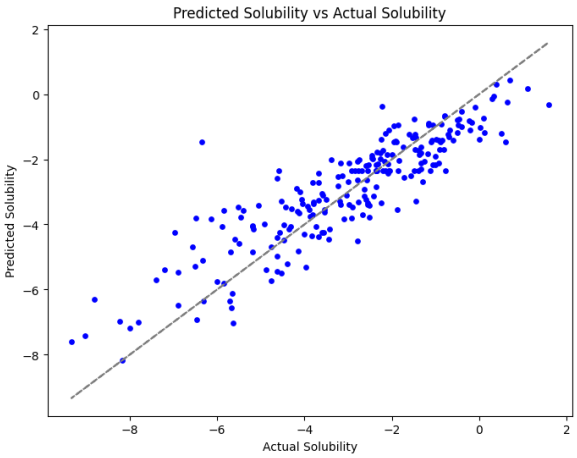


Figure 6: Ground truth vs predicted solubility values

However, it failed to produce better results than the original paper *ESOL: Estimating Aqueous Solubility Directly from Molecular Structure*. Indeed, the average MSE loss for the ESOL model was 0.822 and the average MAE loss was 0.694. Figure 7 shows the predicted solubility values versus the actual solubility values for Delaney's method.

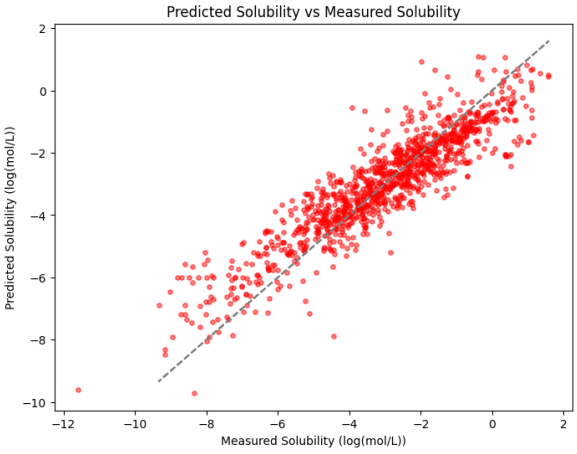


Figure 7: Ground truth vs predicted solubility values with the ESOL model

5 Conclusion

This study demonstrates the effectiveness of representing molecules as graphs and using Graph Neural Networks for molecular property

prediction.
The difference in results between the method from the original paper and this study can probably be explained by the size of the dataset. Indeed, classical Machine Learning methods with hand-crafted features don't require a dataset as big as Deep Learning methods. The lack of big chemical datasets is a well-known issue in cheminformatics and explains that classical Machine Learning methods are still widely used, even though they might produce poorer results.

While this method did not produce better results than classical methods, its strength remains in the fact that it is data-driven and can automatically learn features directly from raw inputs. Therefore, this model can easily be adapted to predict other molecular properties, whereas the ESOL method described in Delaney's paper is limited to predicting aqueous solubility.

6 Implementation

The implementation of this work can be found on *Google Colab*.

References

[1] Sun, M., Zhao, S., Gilvary, C., Elemento, O., Zhou, J., & Wang, F. (2020). *Graph convolutional networks for computational drug development and discovery*. Briefings in Bioinformatics, 21(3), 919-935.

[2] D.Duvenaud, D. Maclaurin et al. (2015). *Convolutional Networks on Graphs for Learning Molecular Fingerprints*.

[3] Wengong Jin, Manolis Kellis. (2021, April 21). *AI for Drug Design - Lecture 16 - Deep Learning in the Life Sciences* [Video]. YouTube.

[4] J. S. Delaney. (2004). *ESOL: Estimating Aqueous Solubility Directly from Molecular Structure*. Journal of Chemical Information and Computer Sciences, 44(3), 1000–1005.

[5] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner and G. Monfardini. (2009, January 1). *The Graph Neural Network Model*. IEEE Transactions on Neural Networks, vol. 20, no. 1.

[6] Abid Ali Awan. (2022, July 21). *A Comprehensive Introduction to Graph Neural Networks (GNNs)*.

[7] Daisy Sharma, Mohit Soni, Sandeep Kumar, GD Gupta. (April.-June.2009). *Solubility Enhancement – Eminent Role in Poorly Soluble Drugs*. Research J. Pharm. and Tech.2(2), Page 220-224.

[8] E. Vitz, J. W. Moore and al. (2023, July 15). 10.19: *Solubility and Molecular structure*. Chemical Principles through Integrated Multiple Exemplars.

[9] Catherine E. Housecroft and Alan G. Sharpe. (2005). *Inorganic Chemistry* (2nd ed.). Pearson Prentice-Hal. p. 100. ISBN 0130-39913-2.

[10] Paul Von Ragué Schleyer. (2001). *Introduction: Aromaticity*. Chemical Reviews, 101(5), 1115–1118.

[11] Clayden, J.; Greeves, N. and Warren, S. (2012) *Organic Chemistry*. Oxford University Press. p.161, p. 145., p.302 ISBN 0-19-927029-5

[12] D. Weininger, (1988). *SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules*. Journal of Chemical Information and Computer Sciences.

[13] T. N. Kipf, and M. Welling. (2016, September 9). *Semi-Supervised Classification with Graph Convolutional Networks*.

[14] W. Ahmad, H. Tayara, and K. T. Chong. (2023). *Attention-Based Graph Neural Network for molecular solubility prediction*. ACS Omega, 8(3), 3236–3244.