Molecule2Vec: Vector Space Representation of Organic Molecules for prediction of properties using Deep Neural networks

Avijit Ghosh, Debasis Sarkar.

Indian Institute of Technology, Kharagpur.

7th EuCheMS Chemistry Congress

30th August, 2018

Introduction

- Solubility is impacted by various factors
- Dipole - induced dipole interactions, Dipole-dipole attractions, Hydrogen bonding
- Solubility of Hydrocarbons are of special interest in the fields of drug delivery, dyeing, paints, and anywhere that an aqueous solution is desired.

Research Problems

- Aqueous solubility of organic molecules is a problem in research applications where new compounds need to by synthesized and the solubility is paramount.
- It is, therefore, of special research interest to come up with a robust method to be able to predict the solubility of hydrocarbons.
- Generalize this method to be able to predict the physical properties of unsynthesized compounds using Machine learning techniques.

Motivation

- What if there was a way to get the solubility of a compound without even synthesizing them?
- Can we automate the process of discovering features from synthetically designed molecules?

Idea

Structure of Synthetic Molecules

Each molecule's 3D structure can be represented, in a standardized manner, as a text document called the **Structure Data File(SDF).**

Vector Space Models

A **Vector Space Model** is an algebraic model used to represent text documents, and finds wide applications in the field of Natural Language Processing and Information Retrieval, where entire text documents are processed, and represented as a vector.

Automatic feature discovery

Combining these two pieces of information we can **automatically discover features** from SDF files and use them to design a model to predict solubility.

Structure Data Files - SDF

```
12 12  0  0  0  0  0  0  0  0999 V2000
    3.2917    3.3940    0.2349 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.9023    3.5389    0.2241 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.8618    2.1207    0.1613 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.0830    2.4105    0.1396 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.0425    0.9922    0.0768 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.6531    1.1372    0.0660 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.9292    4.2719    0.3006 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.4588    4.5296    0.2813 H   0  0  0  0  0  0  0  0  0  0  0  0
    4.9429    2.0079    0.1699 H   0  0  0  0  0  0  0  0  0  0  0  0
    0.0019    2.5232    0.1310 H   0  0  0  0  0  0  0  0  0  0  0  0
    3.4861    0.0016    0.0196 H   0  0  0  0  0  0  0  0  0  0  0  0
    1.0156    0.2592    0.0002 H   0  0  0  0  0  0  0  0  0  0  0  0
  1  3  1  0  0  0  0
  1  7  1  0  0  0  0
  2  1  2  0  0  0  0
  2  8  1  0  0  0  0
  3  5  2  0  0  0  0
  3  9  1  0  0  0  0
  4  2  1  0  0  0  0
  4 10  1  0  0  0  0
```

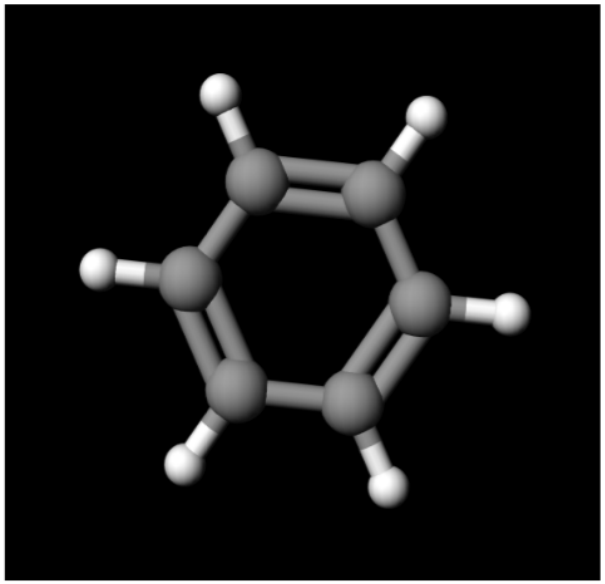


Figure 1: Source: NIST.gov

Word2Vec

- Word2vec is a method of computing vector representations of words introduced by a team of researchers at Google by Mikolov et al. (2013).
- It is a two-layer neural net that processes text. Its input is a text corpus and its output is a set of vectors: feature vectors for words in that corpus.
- While Word2vec is not a deep neural network, it turns text into a numerical form that deep nets can understand.

Word2Vec

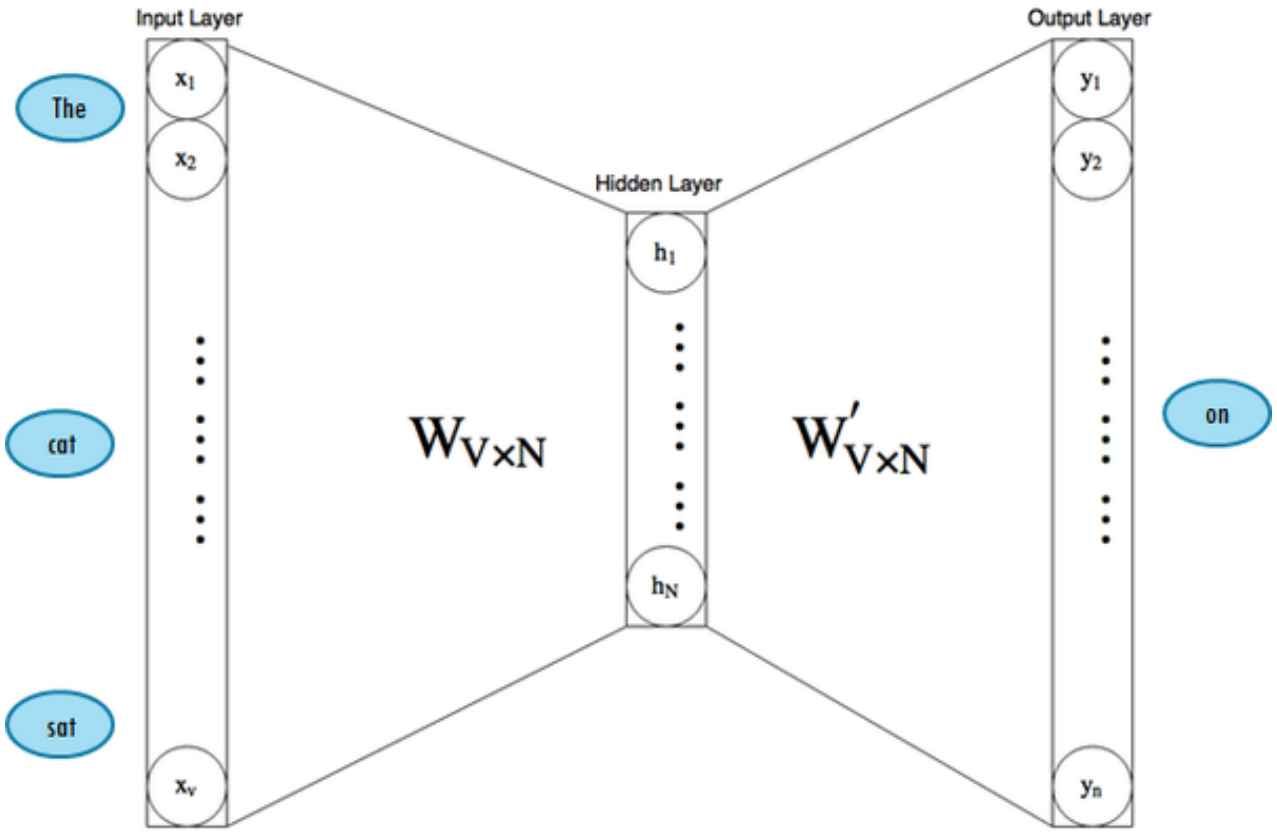


Figure 2: Source: Quora

Doc2Vec

- Doc2Vec, derived from Wor2Vec, was first discussed by Le and Mikolov (2014). It is a method to represent entire complex documents as uniformly sized vectors, regardless of their length.
- To achieve this, the base model of Word2Vec was enhanced by adding a paragraph ID vector for additional context.
- The resulting final vector representation can represent entire documents.

Doc2Vec

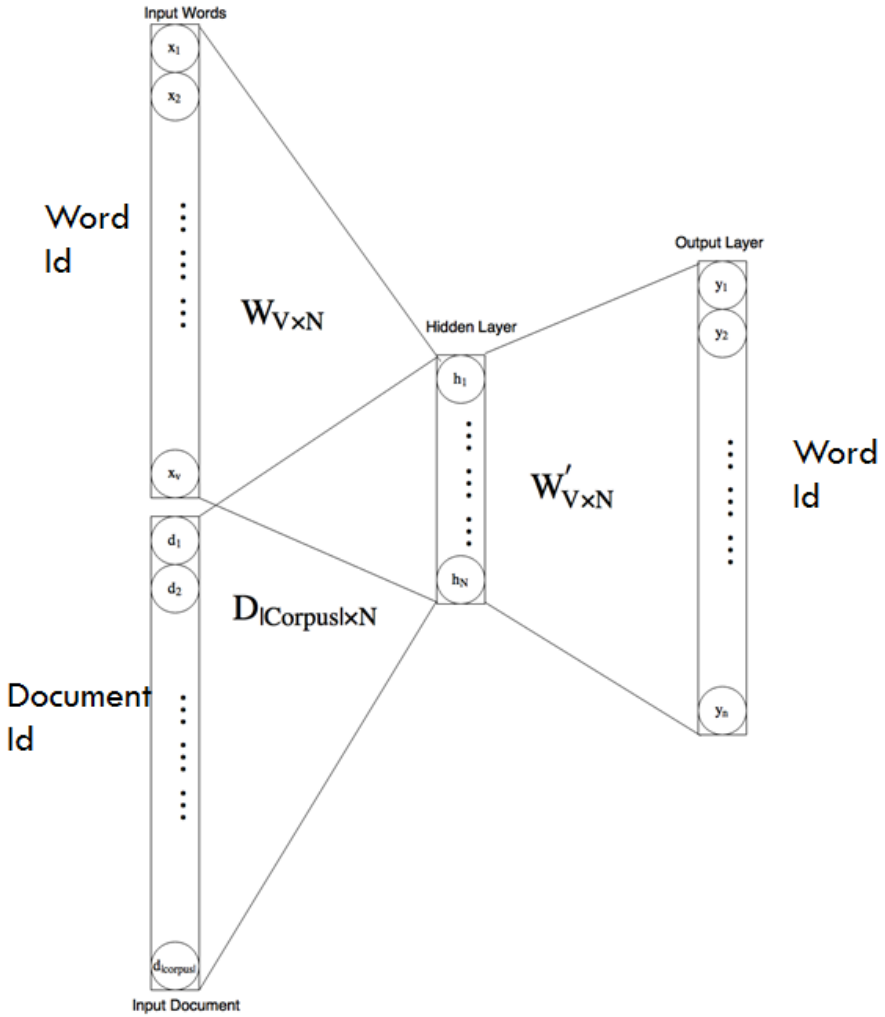


Figure 3: Source: Quora

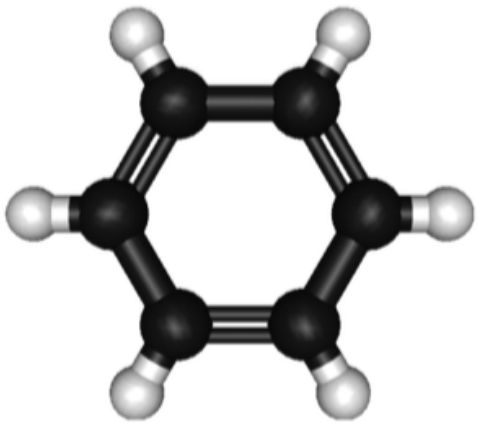
Methodology

I. Data Collection

- Handbook of Aqueous Solubility Data by Yalkowsky et al. (2016).
- The handbook contains the aqueous solubility details of 4661 compounds, which were manually collected and collated in a single database.
- Chemical Structure Data Files (SDF) were collected for each compound from the NIST Chemistry Webbook. Some structures were missing and those compounds were ignored.
- After the complete scraping operation, we had a database consisting of 3263 compounds, with their structures, molecular weight and solubility.

Methodology

II. Converting Structures to Vectors



3D SDF of Molecule

Doc2Vec Training

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ \vdots \\ X_{97} \\ X_{98} \\ X_{99} \\ X_{100} \end{bmatrix}$$

Feature Vector

Methodology

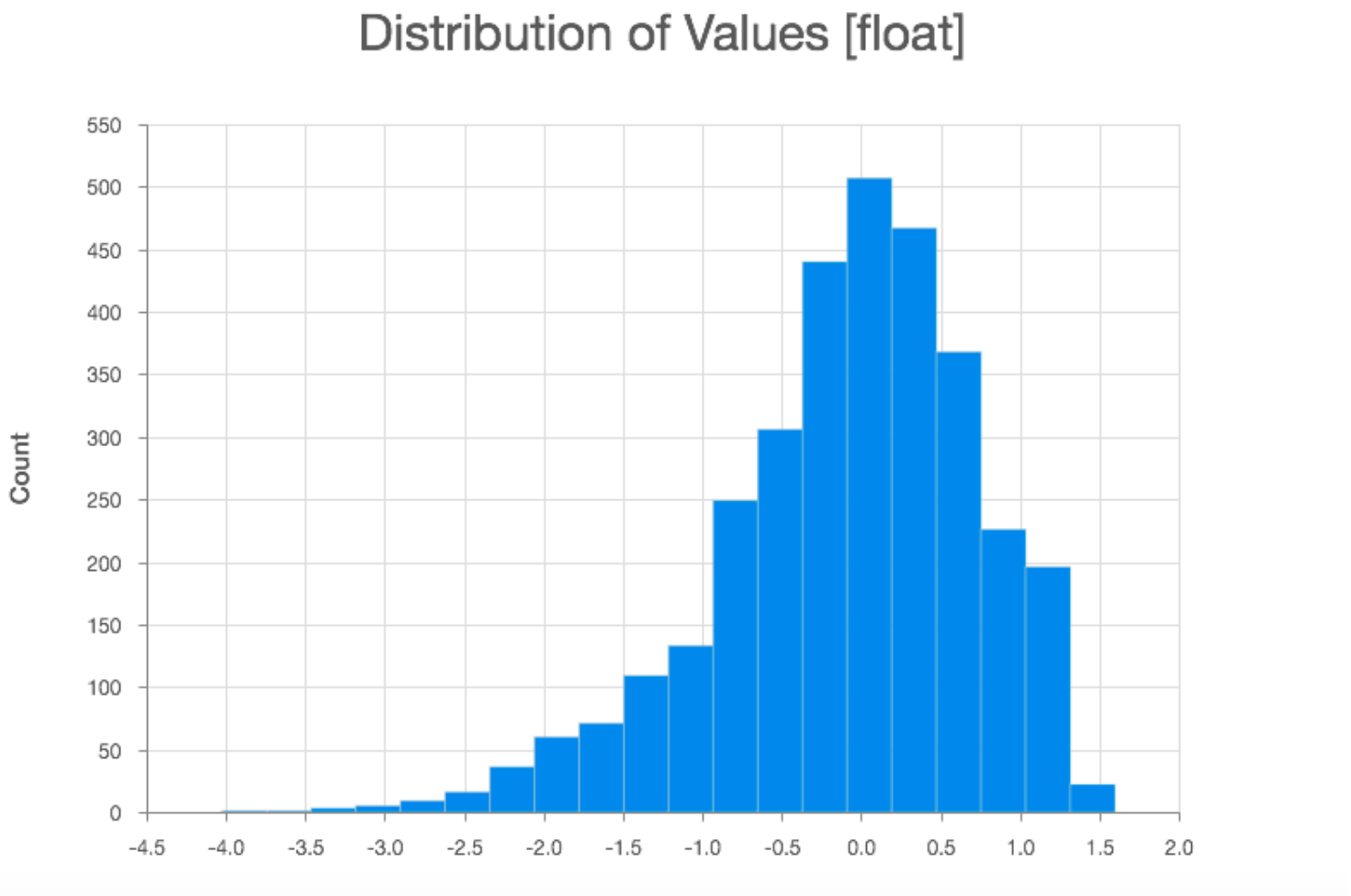
Trained Unsupervised Model

The top five 'similar' compounds to Benzene:

Compound	**Cosine Similarity**
Cyclohexene	0.8457
Tetracene	0.8299
Cumene	0.8229
Chlorobenzene	0.8092
1-Methylphenanthrene	0.8067

Table 1: Benzene Similarity

III. Data Scaling and Normalization



Prediction Algorithms

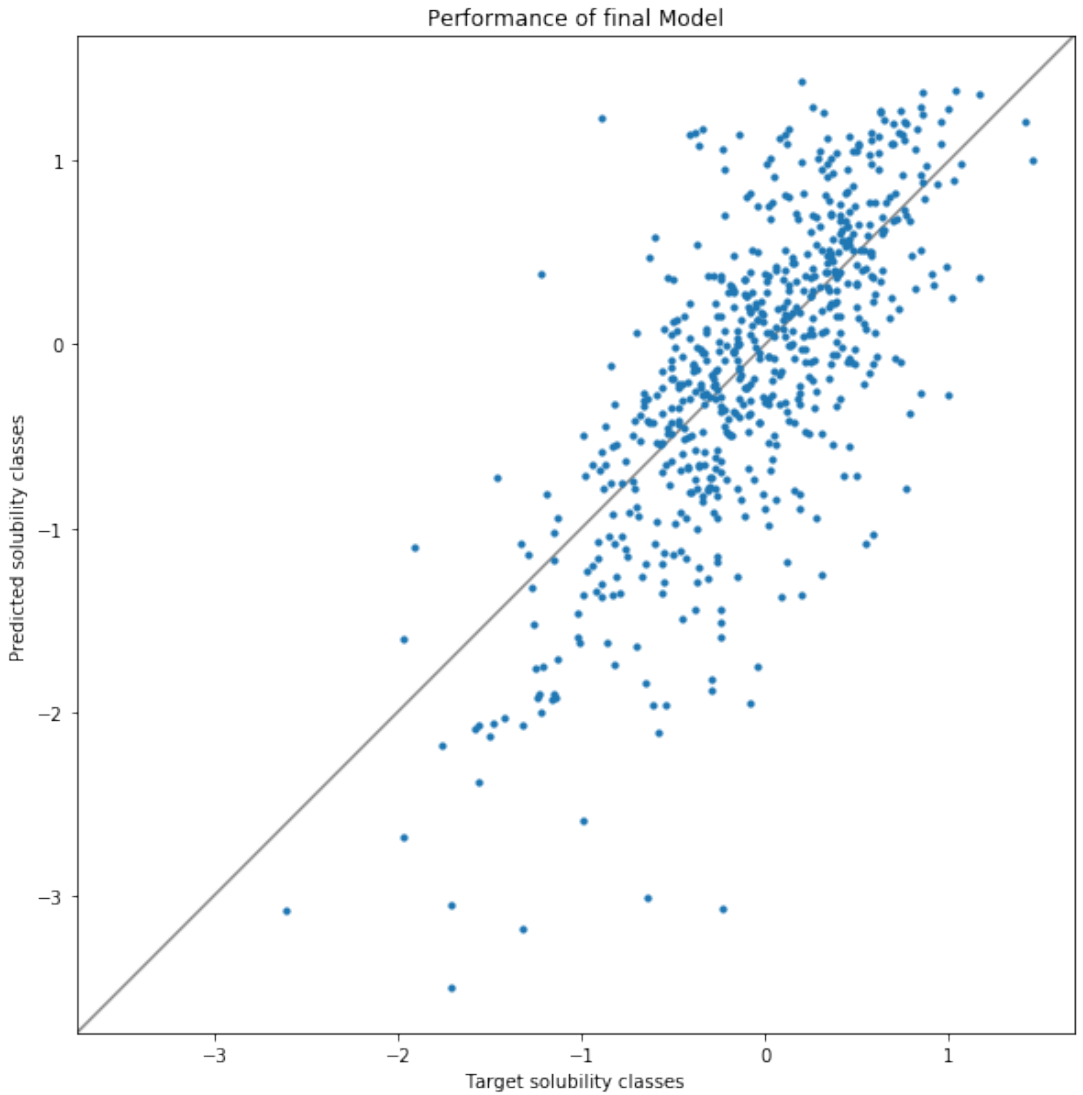
Machine Learning - Simple Regression

- Multiple Linear Regression
- Boosted Trees Regression
- Autosklearn Regression - using combination of various regressors like OLS, XGboost, Random Forest, etc.

Deep Learning - Neural Nets

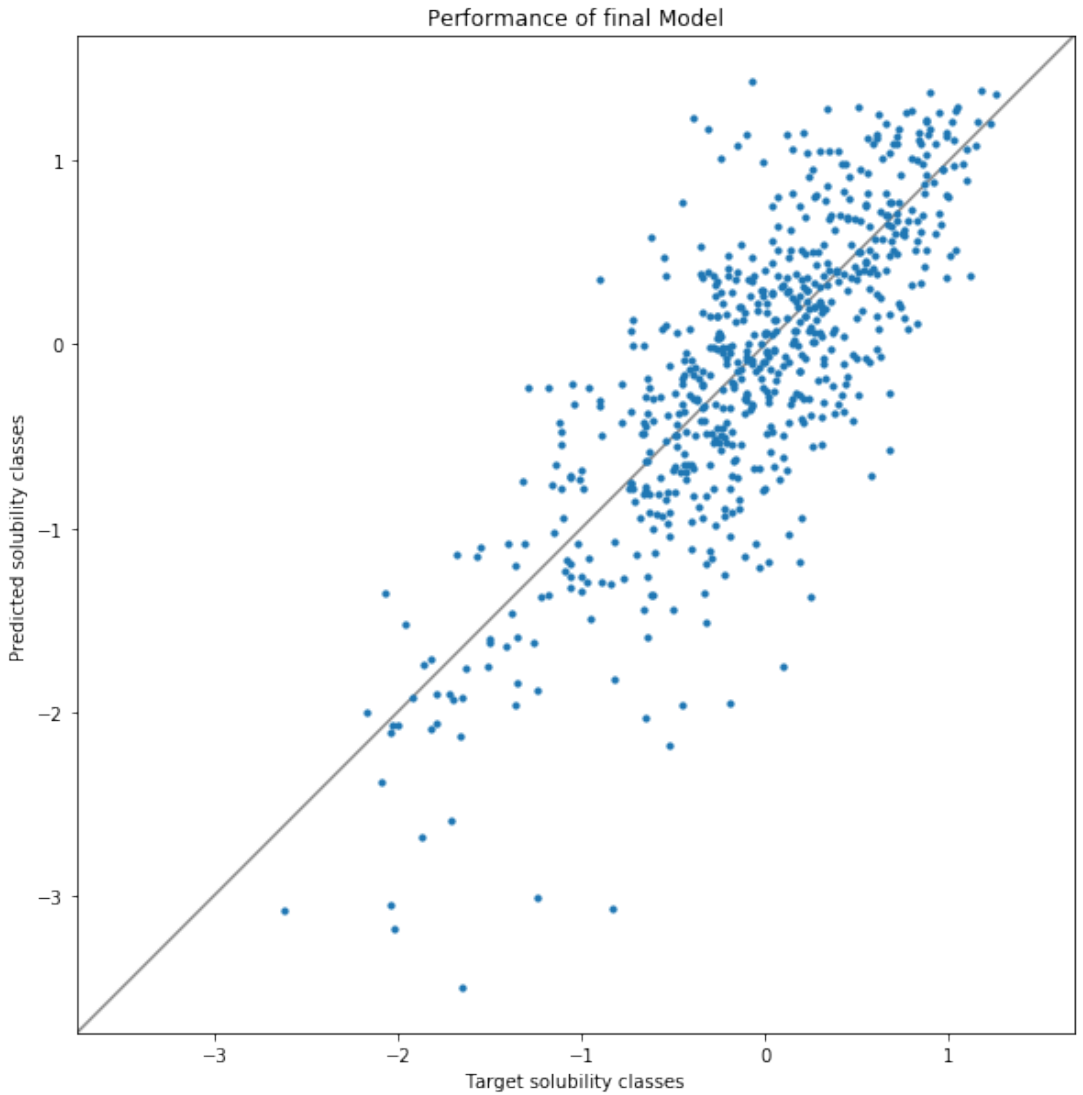
- Dense Neural Network
- Convolutional Neural Network
- SKNN - Automatic Hyperparameter Adjustment

Regression: Linear Regression



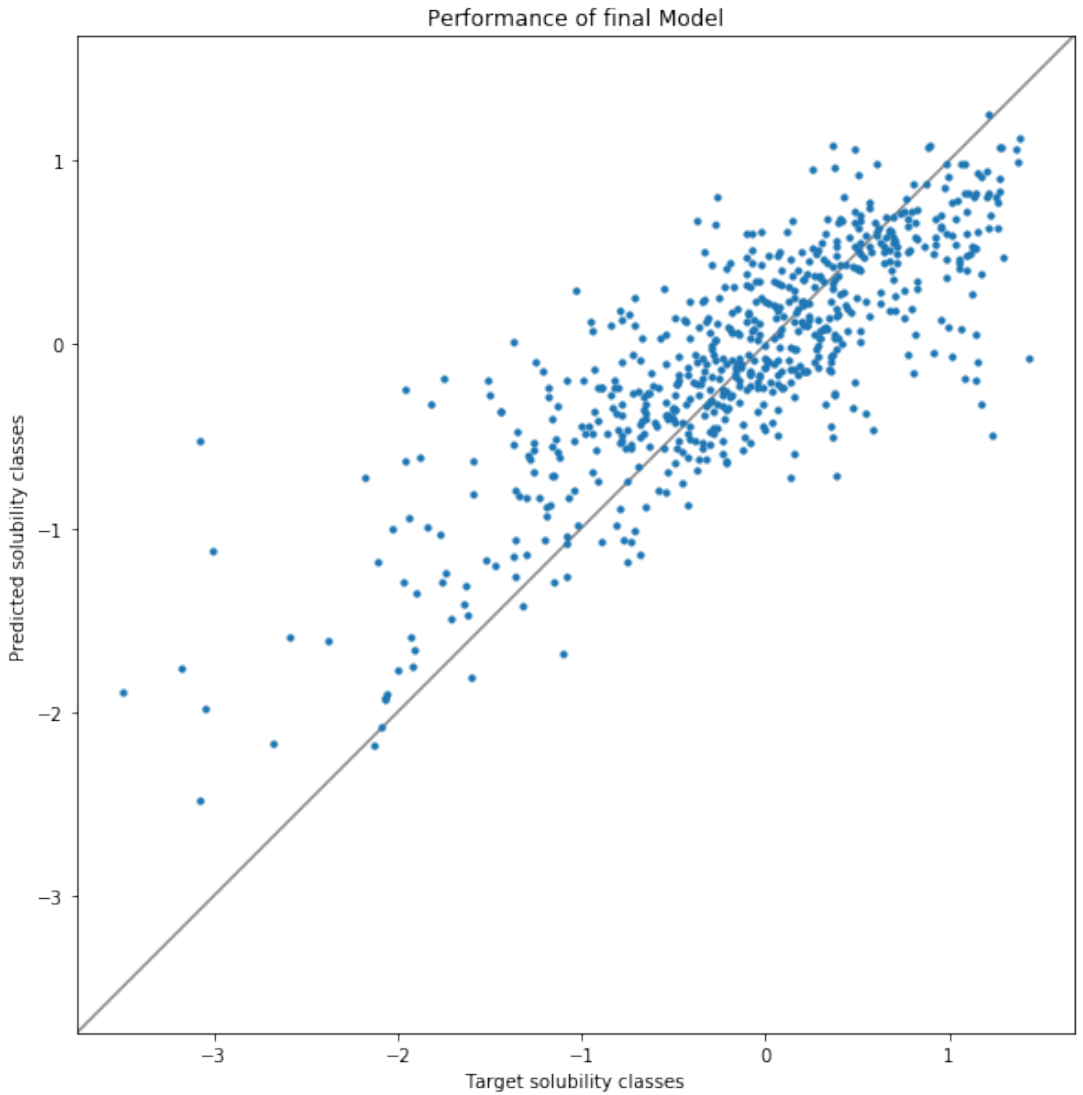
R^2 : 0.5252
MAE: 0.3545 g/litre
RMSE: 1.3486 g/litre

Regression: Boosted Trees Regression



R^2 : 0.6381
MAE: 0.2955 g/litre
RMSE: 1.0847 g/litre

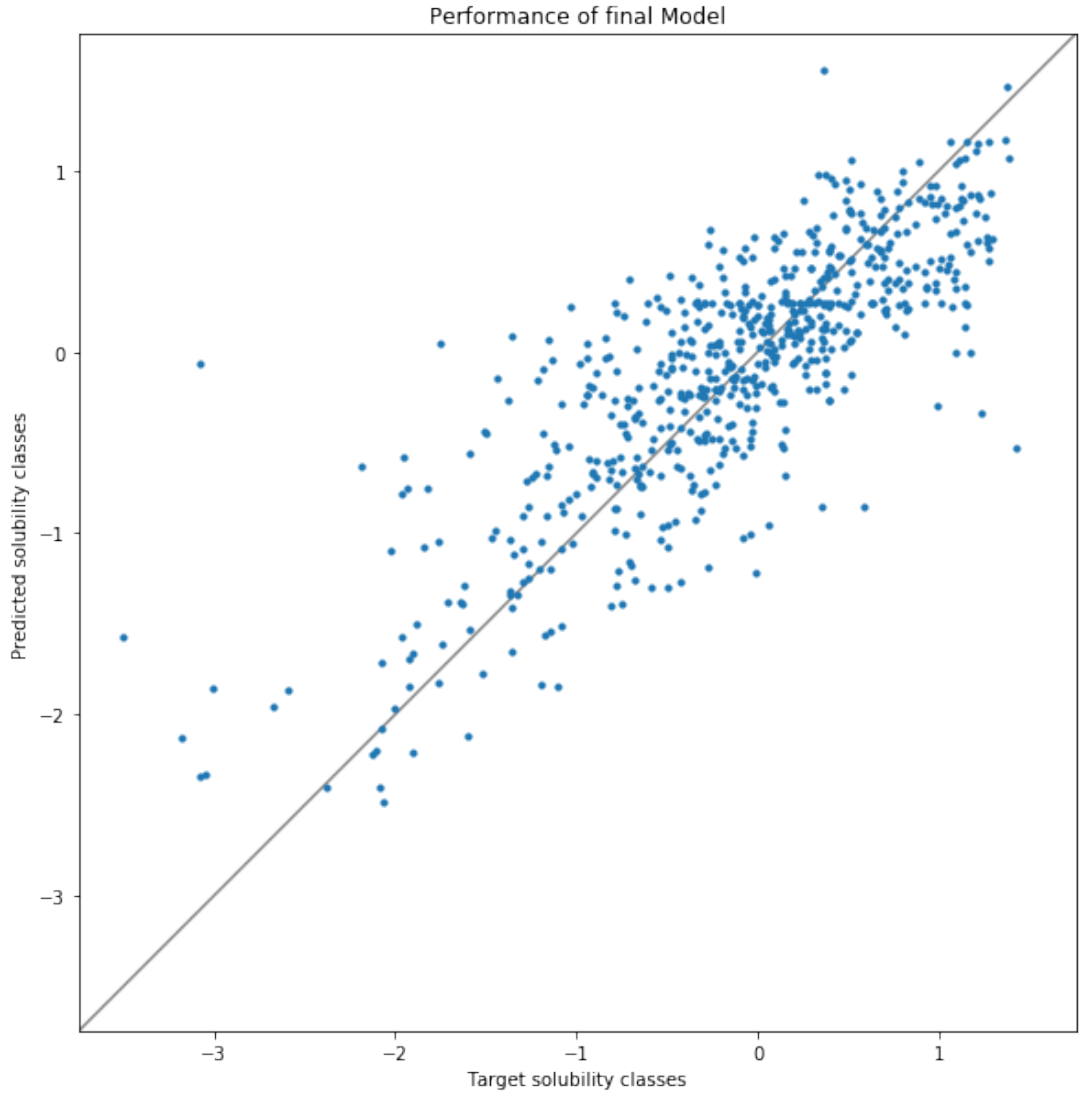
Regression: AutoSklearn Ensemble Regression



R^2 : 0.6540
MAE: 0.3061 g/litre
RMSE: 1.1131 g/litre

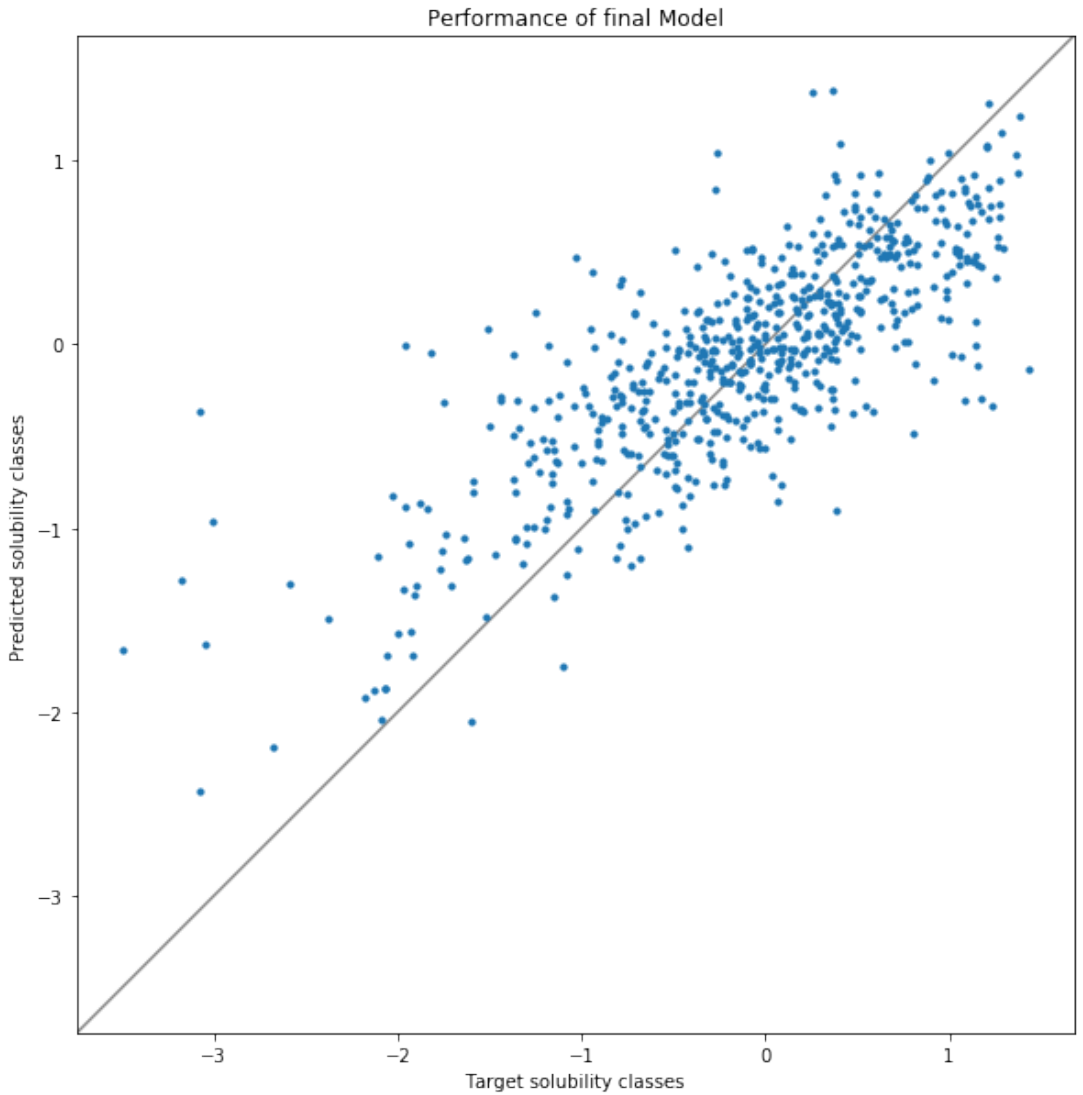
Deep Learning: Dense Neural Network

Layers: $101 \rightarrow 200 \rightarrow 50 \rightarrow 10 \rightarrow 5 \rightarrow 1$



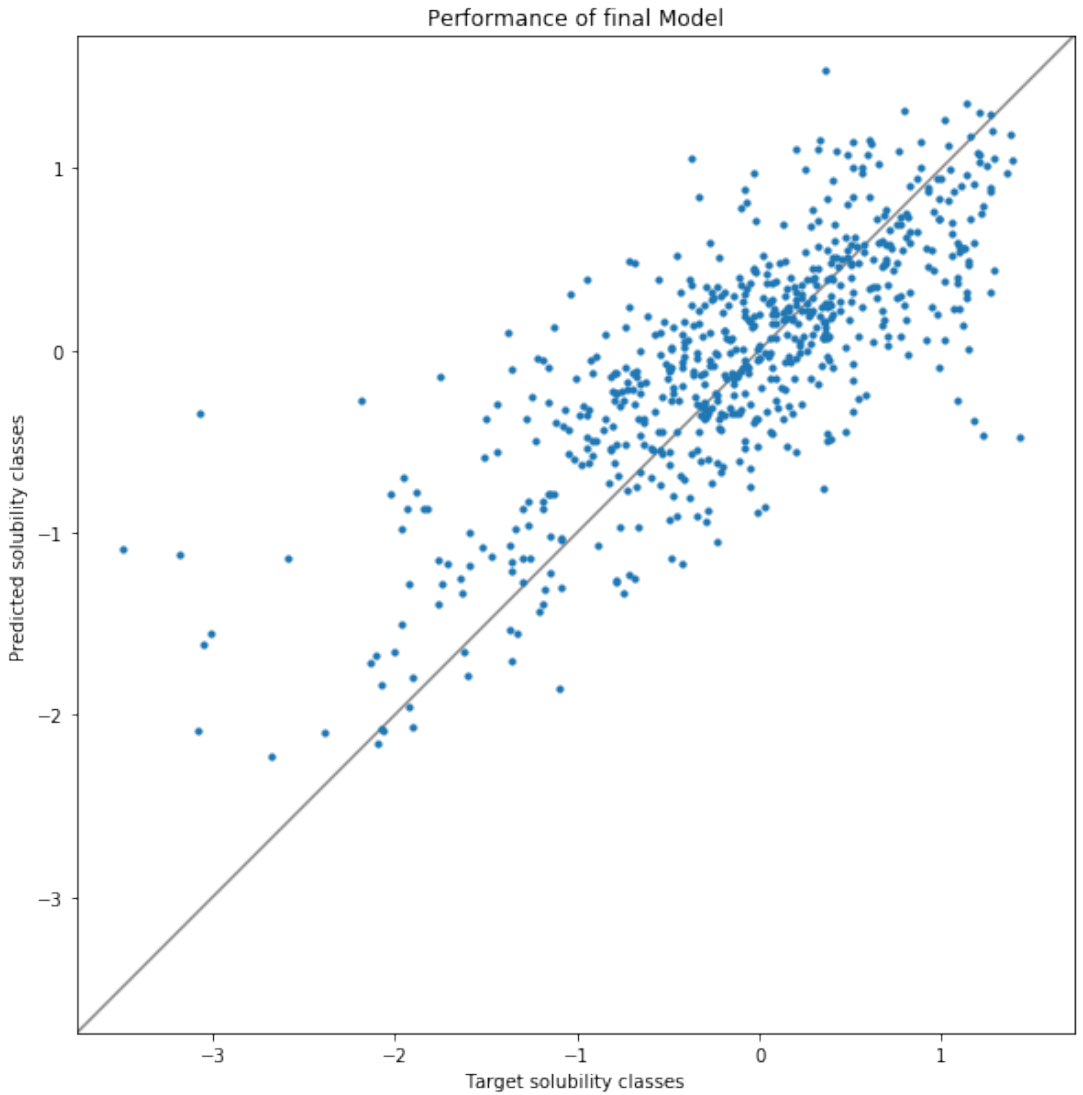
R^2 : 0.6453
MAE: 0.3783 g/litre
RMSE: 1.3804 g/litre

Deep Learning: Convolutional Neural Network



R^2 : 0.6282
MAE: 0.3070 g/litre
RMSE: 1.1441 g/litre

Deep Learning: SkNN - Auto Neural Network



R^2 : 0.5980
MAE: 0.3874 g/litre
RMSE: 1.5594 g/litre

Evaluation results

Method	R^2	MAE	RMSE
Linear Regression	0.525	0.354	1.348
Boosted Trees Regression	0.638	**0.295**	**1.084**
AutoSklearn Ensemble	**0.654**	0.306	1.131
Dense Network	**0.645**	0.378	1.380
Convolutional Network	0.628	**0.307**	**1.144**
SkNN Auto Network	0.598	0.387	1.559

Table 2: Comparison of the Evaluation Performance of the various methods.

Conclusion

- Upto 65% of the solubility patterns can be understood from the structure itself.
- The best model among the regressors is Boosted Trees Regression, which performs slightly better than the deep learning models.
- Probably because different compounds in different solubility zones benefit from the branching of the small zones of binary classifications.
- Deep Learning techniques have room for improvement.

Thanks

Dr. Debasis Sarkar

Volunteering students at the KOSS Winter of Code in the DeepChem project

Fin.