# CH5650: Molecular Data Science and Informatics
## Term Project Report

*Shrivarshan K, MM20B058*

## Introduction:

In this work the review and regeneration of some results are done from the paper "Combining Group-Contribution Concept and Graph Neural Networks Toward Interpretable Molecular Property Models" by Adem R. N. Aouichaoui, Fan Fan, Seyed Soheil Mansouri, Jens Abildskov, and Gürkan Sin. In their work they attempted to develop two interpretable Graph Neural Networks (GNNs) named as GroupGAT and Attentive Group Contribution Model (AGC) by combining the basic working of few benchmark GNNs such as AFP, FraGAT and the Group Contribution concept. The newly developed GNNs are used to predict the property values of various molecules such as aqueous solubility, melting temperature, enthalpy of fusion, enthalpy of combustion and enthalpy of formation using their structure information in the form of smiles string and the respective property values of the groups/atoms present in the molecule as input. As a part of regeneration of results the working of AGC model is reviewed and compared with the results generated using the model in the paper.

## Datasets:

The presented models are tested on a variety of property data:

- Two publicly available data sets consisting of the Delaney aqueous solubility data of small organic compounds (ESOL) and double plus good (highly curated and validated) melting point data set.
- Three proprietary data sets collected from the database constructed by the Design Institute for Physical Properties under the American Institute of Chemical Engineers (AIChE DIPPR) consisting of the enthalpies of formation, fusion, and combustion.

A general description of the data set can be found in Table 1. Summary statistics of the data can be found in Table 2, while the distribution of the data can be seen in Figure 1.
An overview of the various type of molecules included in the study can be seen in Table 3.

| Data set | Symbol | Unit | Size | Description |
|---|---|---|---|---|
| Delaney small aqueous solubility data set (ESOL) | Aq. sol. | $Log_{10}$ (mol/L) | 1128 | Water solubility data for small organic molecules at ambient conditions |
| Bradley double plus good melting points | $T_m$ | °C | 3035 | Temperature at which a compound transitions from the solid to the liquid phase at 1 atm |
| Enthalpy of formation | $\Delta H_{for}$ | kJ/kmol | 741 | hange in enthalpy associated with the formation reaction of the compound in the ideal gas state from its constituent elements at the standard state at 298.15 K and 1 atm |
| Enthalpy of fusion | $\Delta H_{fus}$ | kJ/kmol | 730 | hange in molar enthalpy associated with the isothermal transition of a solid into liquid at its melting point and 1 atm |
| Enthalpy of combustion | $\Delta H_{comb}$ | kJ/kmol | 847 | Increase in enthalpy when a compound undergoes oxidation at 298.15 K and 1 atm |

Table 1: Description of Datasets used

| Data set | Unit | Mean | Std | Min | 25% | 50% | 75% | Max |
|---|---|---|---|---|---|---|---|---|
| Aq. sol. | $Log_{10}$ (mol/L) | −3.02 | 2.10 | −11.60 | −4.28 | −2.80 | −1.58 | 1.58 |
| $T_m$ | °C | 63.10 | 95.88 | −188.00 | 5.00 | 64.00 | 130.00 | 438.00 |
| $\Delta H_{for}$ | kJ/mol | −190.95 | 306.98 | −3385.40 | −316.83 | −150.63 | −14.41 | 397.80 |
| $\Delta H_{fus}$ | kJ/mol | 17.54 | 18.57 | 0.02 | 7.90 | 12.32 | 20.75 | 195.77 |
| $\Delta H_{comb}$ | kJ/mol | −4304.99 | 3107.97 | −40000.00 | −5240.00 | −3730.00 | −2550.00 | −164.00 |

Table 2: Summary Statistics of used Data sets

| Class | Aq. sol. | $T_m$ | $\Delta H_{for}$ | $\Delta H_{fus}$ | $\Delta H_{comb}$ |
|---|---|---|---|---|---|
| Hydrocarbons | 155 | 325 | 224 | 253 | 288 |
| Oxygenated | 285 | 777 | 199 | 212 | 290 |
| Nitrogenated | 48 | 233 | 60 | 56 | 85 |
| Chlorinated | 92 | 87 | 40 | 30 | 11 |
| Fluorinated | 4 | 23 | 24 | 18 | 4 |
| Brominated | 26 | 73 | 11 | 13 | 6 |
| Iodinated | 9 | 21 | 5 | 2 | 1 |
| Phosphorus containing | 0 | 5 | 1 | 1 | 0 |
| Sulfonated | 12 | 44 | 41 | 43 | 42 |
| Silicon containing | 0 | 6 | 0 | 0 | 0 |
| Multifunctional | 497 | 1441 | 116 | 102 | 120 |
| Total | 1128 | 3035 | 741 | 730 | 847 |

Table 3: Chemical Diversity Analysis of Data Set Used

As a part of this work, ESOL dataset considered to test the predictive power of developed AGC model. The dataset contains the name of each molecule, smiles string and the aqueous solubility corresponding to each molecule. A part of dataset and distribution of the dataset is shown in Figure 1 and 2.

| | Name | Const_Value | ESOL_Value | SMILES |
|---|---|---|---|---|
| 0 | Amigdalin | -0.77 | -0.974 | OCC3OC(OCC2OC(OC(C#N)c1ccccc1)C(O)C(O)C2O)C(O)... |
| 1 | Fenfuram | -3.30 | -2.885 | Cc1occc1C(=O)Nc2ccccc2 |
| 2 | citral | -2.06 | -2.579 | CC(C)=CCCC(C)=CC(=O) |
| 3 | Picene | -7.87 | -6.618 | c1ccc2c(c1)ccc3c2ccc4c5ccccc5ccc43 |
| 4 | Thiophene | -1.33 | -2.232 | c1ccsc1 |
| 5 | benzothiazole | -1.50 | -2.733 | c2ccc1scnc1c2 |
| 6 | 2,2,4,6,6'-PCB | -7.32 | -6.545 | Clc1cc(Cl)c(c(Cl)c1)c2c(Cl)cccc2Cl |
| 7 | Estradiol | -5.03 | -4.138 | CC12CCC3C(CCc4cc(O)ccc34)C2CCC1O |
| 8 | Dieldrin | -6.29 | -4.533 | ClC4=C(Cl)C5(Cl)C3C1CC(C2OC12)C3C4(Cl)C5(Cl)Cl |
| 9 | Rotenone | -4.42 | -5.246 | COc5cc4OCC3Oc2c1CC(Oc1ccc2C(=O)C3c4cc5OC)C(C)=C |

Figure 1: First 10 datapoints of ESOL dataset

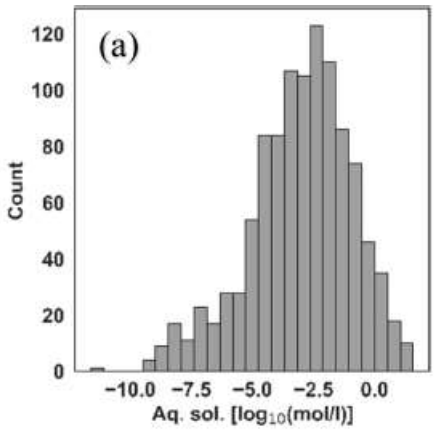


Figure 2: Distribution of ESOL dataset

## Methods and Procedures:

### Molecular Graph Features:

The molecule with N atoms and M bonds is considered as a graph G = {V, E, $X_{atom}$, $X_{bond}$} where V = {$v_1$, $v_2$,⋯,$v_N$} is a set of vertices or atoms, E = {$e_1$, $e_2$,⋯, $e_M$} are the edges or bonds connecting the atoms. $X_{atom}$={$x_1^{atom}$, $x_2^{atom}$, ..., $x_N^{atom}$} is the atom feature matrix with dimension of N×$F_v$ where $F_v$ denotes the size of the atom features. $X_{bond}$={$x_1^{bond}$, $x_2^{bond}$, ..., $x_M^{bond}$} is the bond feature matrix with dimension of M×$F_e$ where $F_e$ denotes the size of the bond features. The features are listed in the Figure 3.

**Table 1. Node (Atom) Features**

| Feature | Description | Size |
|---|---|---|
| Atom type | type of atom (C, N, O, S, F, Cl, Br, I, P) (one-hot encoding) | 9 |
| No. of bonds | number of bonds attached to the atom (0, 1, 2, 3, 4) (one-hot encoding) | 5 |
| No. of Hs | number of hydrogens attached to the atom (0, 1, 2, 3, 4) (one-hot encoding) | 5 |
| Explicit valency | explicit valency (0, 1, 2, 3, 4, 5) (one-hot encoding) | 6 |
| Hybridization | hybridization (sp, sp2, sp3, sp3d, sp3d2) (one-hot encoding) | 5 |
| Aromaticity | whether the atom is part of an aromatic system (0, 1) | 1 |
| Chirality center | whether the atom is a center of chirality (0, 1) | 1 |
| Chirality type | type of chirality the atom is involved in (R, S) | 2 |
| Formal charge | charge assigned to individual atoms in a molecule (int) | 1 |

**Table 2. Edge (Bond) Features**

| Feature | Description | Size |
|---|---|---|
| Bond type | bond type (single, double, triple, aromatic) (one-hot encoding) | 4 |
| Conjugation | whether the bond is conjugated (0, 1) | 1 |
| Ring | whether the bond is part of a ring (0, 1) | 1 |
| Bond stereo | bond stereochemistry (none, any, Z/E, cis/trans) (one-hot encoding) | 6 |

Figure 3: Atom and Bond Features

All these features are generated using RDKit and the molecular graphs are generated using DGL Framework. The relevant codes used for generating features using the rdkit library are available in the "GC-GNN\src\feature".

```python
import numpy as np
from rdkit import Chem
from rdkit.Chem import AllChem, rdMolDescriptors
from rdkit.Chem.EState import EState
import torch

ATOM_VOCAB = ['C', 'N', 'O', 'S', 'F', 'P', 'Cl', 'Br', 'I']


def one_of_k_encoding_unk(x, allowable_set):
    # one-hot features converter
    if x not in allowable_set:
        x = allowable_set[-1]
    return list(map(lambda s: float(x == s), allowable_set))


def chirality_type(x, allowable_set):
    # atom's chirality type
    if x.HasProp('_CIPCode'):
        return one_of_k_encoding_unk(str(x.GetProp('_CIPCode')), allowable_set)
    else:
        return [0, 0]
```

Figure 4: Sample code showing working of featurizer

## Model Building:

### Attentive Group Contribution Model (AGC):

The working of the model is shown in the figure below.

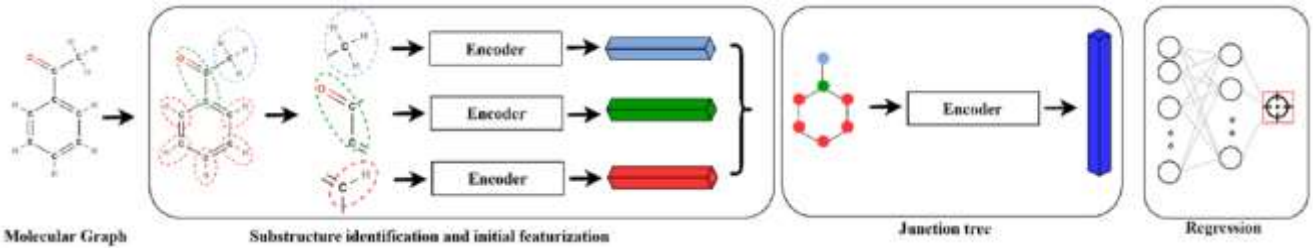


Figure 6: Working of AGC Model

The whole molecular graph is fragmented using the function JT_SubGraph function from "utils\junctiontree_encoder" folder. The scheme used for fragmentation is given in MG_plus_reference dataset. The dataset contains the information of each first order groups with its priority value and SMARTs string. A part of dataset is shown below.

| First-Order Group | Priority | SMARTs |
|---|---|---|
| CH3 | 17 | [CH3;!$(*c)] |
| CH2 | 17 | [CH2;!R;!$(*c)] |
| CH | 17 | [CH1;!R;!$(*c)] |
| C | 17 | [CH0;!R;!$(*c)] |
| CH2=CH | 2 | [CH2;!R]=[CH1;!R;!$(*c)] |
| CH=CH | 2 | [CH1;!R;!$(*c)]=[CH1;!R;!$(*c)] |
| CH2=C | 2 | [CH2;!R]=[CH0;!R;!$(*c)] |
| CH=C | 2 | [CH1;!R]=[CH0;!R] |
| C=C | 2 | [CH0;!R]=[CH0;!R] |
| CH2=C=CH | 1 | [CH2;!R]=[CH0;!R]=[CH1;!R] |
| CH2=C=C | 1 | [CH2;!R]=[CH0;!R]=[CH0;!R] |
| CH=C=CH | 1 | [CH1;!R]=[CH0;!R]=[CH1;!R] |
| CH#C | 1 | [CH1]#[CH0] |
| C#C | 1 | [CH0]#[CH0] |
| ACH | 14 | [cH1] |
| AC fused with aromatic ring | 2 | [cH0;R2;!$(*C)] |
| AC fused with non-aromatic ring | 2 | [cH0;R2;$(*C)] |

Figure 7: Sample data from MG_plus_reference dataset

## Training and optimization:

The data was split randomly with a split ratio of 80%–10%–10% for training, validation, and testing, respectively. The split was the same for each property across the models. The split is done using **Splitter** function from the "utils\splitter" folder.

To avoid the way splitting to influence the results the training procedure is done 500 times each with different split. But due to unavailability of computationally efficient system in this work I ran it 10 times and chose the best model for prediction.

All models were developed using DGL for the graph-learning framework with Pytorch (v1.12.0) as the backend deep-learning framework.

### Hyperparameter tuning:

- The model hyperparameters were determined using Bayesian optimization (BO) by minimizing the validation loss.
- Throughout all training instances, early stopping was employed to retain the best performing models with respect to the validation set and to avoid overfitting.
- The maximum number of epochs was set to 300. The dropout rate, weight decay, and learning rate reduction factors were all considered hyperparameters.
- For the MLP part of the GNN models, only two layers were considered for which the number of neurons in the first layer was considered hyperparameter subjected to the optimization, while the number of neurons in the second layer was considered half of that of the first layer.

The search space used for hyperparameter tuning is given in the following table.

| Model | Hidden dimensions | Initial learning rate | Learning rate reduction factor | Weight decay | Dropout rate | Node-level embedding layers, $T$ | Graph-level embedding, $L$ |
|---|---|---|---|---|---|---|---|
| MPNN | [16, 256] | 1e-[1, 5] | [0.4, 0.95] | [0, 0.1] | [0, 0.5] | int([1, 6]) | – |
| D-MPNN | [16, 256] | 1e-[1, 5] | [0.4, 0.95] | [0, 0.1] | [0, 0.5] | int([1, 6]) | – |
| AFP | [16, 256] | 1e-[1, 5] | [0.4, 0.95] | [0, 0.1] | [0, 0.5] | int([1, 6]) | int([1, 6]) |
| FraGAT | [16, 256] | 1e-[1, 5] | [0.4, 0.95] | [0, 0.1] | [0, 0.9] | int([1, 6]) | int([1, 6]) |
| AGC | [16, 256] | 1e-[1, 5] | [0.4, 0.95] | [0, 0.1] | [0, 0.9] | int([1, 6]) | int([1, 6]) |
| GroupGAT | [16, 256] | 1e-[1, 5] | [0.4, 0.95] | [0, 0.1] | [0, 0.9] | int([1, 6]) | int([1, 6]) |

Table 4: Hyperparameter search space

The code used for hyperparameter tuning is there in GC-GNN\new_frag_optimization.py

Though as a part of this work I skipped the optimization part as the optimized hyperparameters is attached with the paper which is shown below.

*Table S 2: Optimal hyperparameters for various models applied on Aq. sol.*

| Model | Hidden Dimensions | Initial learning rate | Learning rate reduction factor | Weight decay | Dropout rate | Nr. node level embedding layers, T | Nr. of graph level embedding, L |
|---|---|---|---|---|---|---|---|
| **MPNN** | 118 | 1e-3.9 | 0.9 | 0 | 0.00 | 6 | N.A |
| **D-MPNN** | 47 | 1e-2.4 | 0.4 | 0 | 0.25 | 1 | N.A |
| **AFP** | 111 | 1e-2.4 | 0.5 | 0 | 0.15 | 1 | 1 |
| **FraGAT** | 122 | 1e-3.5 | 0.4 | 1e-05 | 0.10 | 1 | 2 |
| **AGC** | 36 | 1e-2.3 | 0.8 | 0 | 0.20 | 1 | 1 |
| **GC-GAT** | 226 | 1e-3.0 | 0.8 | 1e-06 | 0.00 | 1 | 1 |

Table 5: Optimal hyperparameters for ESOL dataset

The hyperparameters given for AGC model is directly used as initial parameters and the model is trained.

## Results and Comparison:

Various metrics are used in the paper to test the predictive power of the model. In this work, MAE, RMSE, R2 score of training, validation, test and overall datasets and parity plot are used as metrics to test the predictive power of the model.

### Predictions:

The predicted aqueous solubility values and actual solubility of few datapoints from train, test and validation dataset are shown below.

| | Target | Predict |
|---|---|---|
| **897** | -1.460 | -1.382201 |
| **898** | -2.360 | -2.493675 |
| **899** | -2.878 | -2.987362 |
| **900** | -3.660 | -3.748746 |
| **901** | -3.953 | -3.863875 |

Prediction on train data

| | Target | Predict |
|---|---|---|
| **107** | -4.150 | -3.773423 |
| **108** | -0.830 | -1.277600 |
| **109** | -3.290 | -3.300147 |
| **110** | -4.800 | -4.484435 |
| **111** | -2.154 | -2.674221 |

Predictions on the validation data

| | Target | Predict |
|---|---|---|
| **109** | -9.150001 | -8.614853 |
| **110** | -5.720000 | -5.082323 |
| **111** | 1.120000 | 1.090296 |
| **112** | -3.760000 | -2.925282 |
| **113** | -6.637000 | -6.274870 |

Predictions on the test data

**MAE, RMSE, R2 score:**

As discussed to avoid the influence of splitting in the model's predictive power the training is done 10 times. The scores of each iteration is given below with the seed value.

| | seed | train_R2 | val_R2 | test_R2 | all_R2 | train_MAE | val_MAE | test_MAE | all_MAE | train_RMSE | val_RMSE | test_RMSE | all_RMSE |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1457 | 0.914794 | 0.873454 | 0.879117 | 0.907464 | 0.46333411 | 0.517151 | 0.6088092 | 0.48338 | 0.609457499 | 0.68707544 | 0.786074088 | 0.6374481 |
| 1 | 134 | 0.922168 | 0.884983 | 0.868492 | 0.912672 | 0.43263254 | 0.528904 | 0.642114 | 0.463362 | 0.581067612 | 0.68668302 | 0.812157669 | 0.6192516 |
| 2 | 1342 | 0.941709 | 0.938861 | 0.887431 | 0.936654 | 0.38043219 | 0.420458 | 0.487184 | 0.395195 | 0.504228799 | 0.55399707 | 0.662696758 | 0.5274132 |
| 3 | 4508 | 0.917156 | 0.914909 | 0.883396 | 0.913714 | 0.45323446 | 0.402382 | 0.531656 | 0.456111 | 0.606515435 | 0.56330999 | 0.725758097 | 0.6155468 |
| 4 | 4314 | 0.941019 | 0.897438 | 0.899673 | 0.931604 | 0.37494072 | 0.492509 | 0.5017156 | 0.399427 | 0.501065995 | 0.6674406 | 0.740003082 | 0.5480324 |
| 5 | 2055 | 0.926217 | 0.918604 | 0.904964 | 0.923159 | 0.42031035 | 0.49805 | 0.5164942 | 0.43775 | 0.558248907 | 0.66365633 | 0.663477988 | 0.5808814 |
| 6 | 2504 | 0.943149 | 0.88636 | 0.902459 | 0.934917 | 0.37666598 | 0.502156 | 0.4603339 | 0.397582 | 0.507640575 | 0.66109601 | 0.599552289 | 0.5345945 |
| 7 | 2832 | 0.932613 | 0.861907 | 0.830164 | 0.917083 | 0.42010337 | 0.582419 | 0.6251151 | 0.456939 | 0.550093309 | 0.74891533 | 0.810801712 | 0.6034101 |
| 8 | 3581 | 0.965363 | 0.917602 | 0.922253 | 0.954866 | 0.2792733 | 0.485452 | 0.4909396 | 0.321137 | 0.380120563 | 0.62380746 | 0.659909038 | 0.4451864 |
| 9 | 731 | 0.942649 | 0.893729 | 0.893737 | 0.933627 | 0.38558283 | 0.462116 | 0.5632725 | 0.41114 | 0.503237406 | 0.61812448 | 0.710453581 | 0.5398673 |

Table 6: Metrics of AGC model on ESOL dataset

The model with **maximum R2 score** is with seed value = 3581 with **test_R2 = 0.922** and **overall_R2 = 0.955.**

The **average R2 score** of all the 10 models is **test_R2 = 0.887 and overall_R2 = 0.926.**

The metrics of the model developed in the paper is shown below.

| | | | MAE | | RMSE | | $R_2$ | |
|---|---|---|---|---|---|---|---|---|
| Data set | Model | Nr. parameters | Test | Overall | Test | Overall | Test | Overall |
| Aq. sol. | GC[13] | 70 | – | 0.73 | – | – | – | 0.78 |
| | AGC | 60,017 | 0.44 | 0.39 | 0.61 | 0.53 | 0.92 | 0.94 |
| | GroupGAT | 3,185,025 | 0.36 | 0.32 | 0.51 | 0.43 | 0.94 | 0.96 |

Table 7: Metrics of AGC Model developed in the paper

The model developed by the author gave **test_R2 = 0.92** and **overall_R2 = 0.94**

Though the model is run only for 10 times the R2_scores are pretty close to the R2_scores developed by the author.

**Parity plot:**

The parity plots developed by me and developed on the paper is shown below.

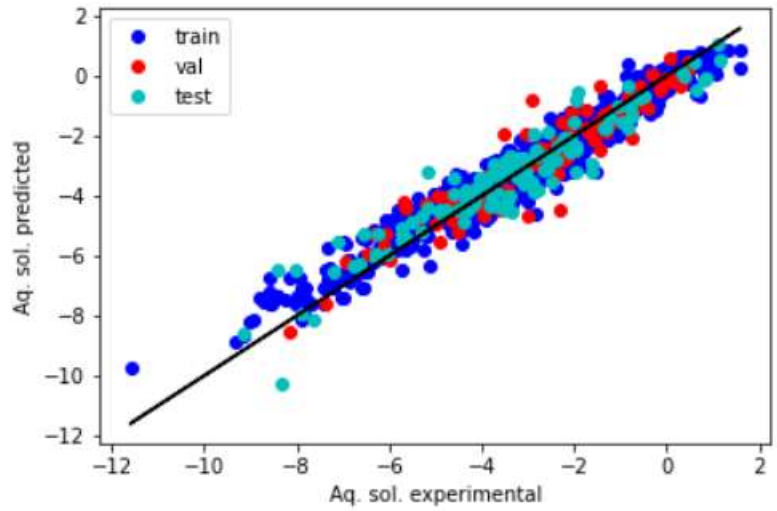


Figure 8: Parity plot developed in this work

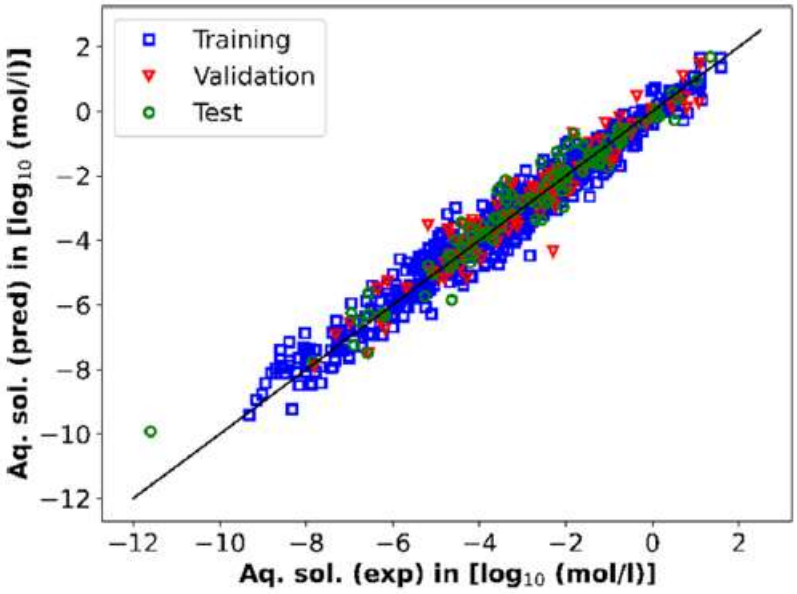


Figure 9: Parity plot developed on the paper

The parity plots look very similar as the R2 scores of the model developed in this work and on the paper are close to each other.

## Conclusion:

Apart from the predictions, there are also attention weights obtained for each molecular fragment which can be viewed as a visualization using the tools in the plottools folder of GC-GNN. All the codes used are from the github folder https://github.com/gsi-lab/GC-GNN. The folder is also cloned into the system and I submit the folder for further verification.

All the other sources used such as MG_plus_reference dataset and tuned hyperparameters are attached with the paper.