This is the results document for the github.com/pcypw2/**GraphSolMol** repository. For any questions regarding the code or results, please contact myself (Phil Wroe) or Joe Redshaw. The contact details for us both are found below:

* [pcypw2@nottingham.ac.uk](mailto:pcypw2@nottingham.ac.uk)
* [psxjr4@nottingham.ac.uk](mailto:psxjr4@nottingham.ac.uk)

**Molecular Descriptors**

The graph representation was compared against the MACCS key fingerprint and ECFP structural bit descriptors as these are the popular choice due to their high predictive performance20,21 (figure 1). Due to time restraints in the project, the ECFP with a radius of 3 was considered which is believed to offer best results as it is able to capture longer range interactions.

**Modelling Methods**

1. *Multiple Linear Regression*

The *AqSolDB* was passed through the ESOL multiple linear regression method as benchmark for predictive success. This was achieved using scikit-learn.

1. *Support Vector Regression*

SVMs implemented by scikit-learn where used to assess kernel performance. Table 1 highlights the kernel methods used. The vectorial kernel matrices were calculated in-house using scikit-learn, whilst the graphical kernel matrices were computed externally through Grakel before being inputted into the SVM.

**Model Building and Evaluation**

The SVM hyperparameters were optimised using a grid search 5-fold cross validation method with an exponentially growing sequence. E.g., c = 2-5, 2-4, …, 25, g=2-15,2-13…, 25, e =1-3,1-2, 12. Prior to model fitting, the data underwent a linear scaler transformation. Both the scaling and grid search optimisation methods were performed using sci-kit learn. The number of iterations (n\_iter) parameter for Weisfeiler -Lehman and Weisfeiler-Lehman Optimal Assignment kernel was optimised by testing a range of n\_iter from 1 to 15 and the model with best performing results was chosen.

Each SVM model was built by with an 80:20 random split the AqSolDBto form a training and test set. The model performance was assessed through calculation of the coefficient of determination (*r2*) (equation 1), which shows the proportion of the solubility that is predictable by the descriptors. The root mean squared error (*RMSE*) (equation 2) was also used to determine how well the predictions fit the data. All analysis was performed with scikit-learn. Where is the predicted and is the experimental solubility value and is the number of points in the dataset.

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |

**Preliminary Analysis**

*Aims of the preliminary analysis:*

* To finetune the AqSolDB, ensuring the correct balance between allowing experimental noise into the model and being as representative of the vast potential of chemical space
* Create a benchmark model to compare support vector regression performance against

The first investigation involved splitting the AqSolDB into three groups to understand if the removal of data with high error had any significant improvement on prediction quality. **AqSolDB-A** contained only group 1 where there was only 1 reported compound and thus no standard deviation. **AqSolDB-B** which had groups 1,2, and 4 which introduced an error of <0.5 SD where there were more than one reported compound across the 9 datasets used to build the AqSolDB. Finally, **AqSolDB-C** which contained all 5 groups which included a standard deviation >0.5 where there are multiple occurrences of a compound.

The ESOL method by Delaney et al. was used to build a normalized MLR model for each of the three datasets (table 1). Whilst AqSolDB-A achieved slight improvement in prediction performance, overall, the three models were largely inaccurate. Given that the AqSolDB was in its raw state and had yet to undergo any pre-processing, this result does not come as a surprise. Figure 1 highlights the extreme ranges of datapoints of lipophilicity (log*P*) and molecular weight (MolWt) that are included in the AqSolDB. In the context of aqueous solubility modelling for a drug discovery project, inclusion of datapoint points over 1000 Da are not relevant and only hinders prediction performance as anything over this MolWt would not be considered in the pharmaceutical industry. A ‘tight’ and ‘broad’ range for MolWt and log*P* were proposed to focus the dataset and make it relevant for drug-like molecules.

**Table 1.** MLR Regression performance of AqSolDB datasets with increasing amount of experimental error. r2 = coefficient of determination and RMSE = root mean squared error.

|  |  |  |
| --- | --- | --- |
|  | *r2* | *RMSE* |
| AqSolDB-A | 0.46 | 1.74 |
| AqSolDB-B | 0.44 | 1.79 |
| AqSolDB-C | 0.44 | 1.77 |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | *MolWt* | *n* | log*P* | Lower *r2 CI* boundary | *r2* | Upper *CI* boundary | *RMSE* |
| *1* | *≤500* | *9076* | *-7 ≤ x ≤ 7* | *0.55* | *0.57* | *0.58* | *1.42* |
| *2* | *≤500* | *9329* | *-11 ≤ x ≤ 11* | *0.58* | *0.59* | *0.60* | *1.46* |
| *3* | *≤1000* | *9391* | *-7 ≤ x ≤ 7* | *0.53* | *0.54* | *0.55* | *1.48* |
| *4* | *≤1000* | *9783* | *-11 ≤ x ≤ 11* | *0.55* | *0.56* | *0.57* | *1.55* |
| *5* | *≤5299* | *9983* | *-41 ≤ x ≤ 68* | *0.43* | *0.44* | *0.46* | *1.77* |

**Table 2.** Results of MLR regression on tight and broad bands of MolWt and logP (tests 1-4) compared against the raw dataset (test 5). n = number of datapoints, CI = confidence internal. Green highlight shows the benchmark model and dataset used for the kernel investigation

**Figure 1.** Histogram plots showing the spread of solubility, molecular weight and lipophilicity for the AqSolDB

When compared to the raw dataset, each test performed significantly better, but there is little difference when comparing between tests. A 95% confidence interval (CI) limit therefore was calculated to determine which dataset, if any, had significantly improvements on prediction quality. Findings from this, show that the dataset 2 had a significant improvement in prediction performance over both dataset 3 and 4, which had a higher cut off for MolWt. However, we decided that inclusion of as many datapoints as possible was the route to go down to increase the flexibility of our models to unseen molecular structures. This increased representation of structures in the extreme ranges of MolWt and logP better depict the chemical space of drug-like molecules better which leads to higher accuracy predictions. If, however, you were predicting a class of small MolWt compounds, using the tighter dataset will yield better results as it is unnecessary to include larger molecular weight compounds. Given the binomial-like distribution of solubility values in figure 1, retaining the extreme solubility datapoints was an import to creating a more generalised model. To confirm this, the MLR equations generated from both test 2 and 4 (equations 1 and 2) were applied to the validation dataset (table 4). Whilst this showed no significant difference between the two as the CI boundaries for each test overlap less error in broader dataset, thereby showing the need of a large representation of different molecules. Therefore we used the models highlighted in green in table 2 and table 3 as our benchmark for test and validation results.

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |

**Table 3.** Applied MLR equation of dataset 2 (equation 1) and dataset 4 (equation 2) on an external validation set

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Equation** | **Lower *r2* CI boundary** | ***r2*** | **Upper *r2* CI boundary** | **RMSE** |
| (1) | 0.47 | 0.51 | 0.54 | 1.41 |
| (2) | 0.48 | 0.52 | 0.56 | 1.33 |

**Kernel Investigation Results**

Vectorial kernels used were linear, radial basis function (RBF), polynomial and sigmoid. As for graph kernels, we explored neighbourhood hashed (NH), random walk (RW), propagation (PG), ordered dag decomposition (ODD), pyramid match (PM), shortest path (SP), graphlet sampling (GS), vertex histogram (VH), Weisfeiler Lehman subtree (WLS) and Weisfeiler-Lehman optimal assignment (WLOA) (table 4).

**Predicted vs Experimental Plots**

Below are the performance plots of each kernel method. The dark blue plot (left) was built from the training dataset, the light blue plot (middle) is from the test set and the green plot (right) is the validation dataset. For polynomial kernel, only the best performing degree is shown (ESOL = 1, MACCS =2, ECFP =2) (figure 2)

**Solubility Challenge Dataset**

Best performing kernel methods were benchmarked using the high accuracy solubility data included in the second solubility challenge. Performance of models from the original paper varied significantly (*r2*=0.09 and *RMSE* =1.14 average results) with the greatest prediction model receiving *r2* = 0.64, and *RMSE* =0.76. Fitting the SC2 dataset on the NH, WLS, WLOA, MACCS RBF and WLOA models obtained impressive results when compared against the average performance of models in the original paper9. The superior performance came from the NH model which had the highest accuracy and shares almost identical model performance against the best performing entry in the original paper. Implying that the graph representation can offer better description of molecular structure against vectorial descriptors that have been used as the as the state-of-the-art. From figure 3, it is clear that more datapoints in the low end of solubility are needed to boost model performance further. Regardless, this clearly demonstrates the potential graph kernels have to aqueous solubility modelling and possibly QSAR modelling of other physiochemical properties.

**Conclusions & Further Work**

SVR models built on structural, physiochemical and graph representations were used to assess vector and graph kernel performance for predicting aqueous solubility. Whilst model performance varied across the graph kernels explored, the neighbourhood hashed, Weisfeiler-Lehman subtree and Weisfeiler-Lehman optimal assignment kernels achieved superior performance compared against the vectorial class. The neighbourhood hashed and Weisfeiler-Lehman subtree kernels offered similar computational requirements as and were generally inexpensive to run, unlike the optimal assignment which was very demanding. Further improvements to enhance the accuracy of the models include creating a dataset which contains a balanced range of molecular structures across the solubility spectrum as a whole. Additionally, careful consideration of the experimental reliability of data prior to model building could result in a lower error overall. Further investigation on graph kernels includes the exploration of indefinite (non-positive definite) graph kernels to extend our understanding of graph representation for aqueous solubility modelling.

Particular work of interest could involve application into different dataset that of high experimental accuracy or focusing on a particular class of compounds. Additionally, many more graph kernel types exist than just the ones mentioned in this research have yet to be applied to predicting solubility, even pilot studies using indefinite kernels could be investigated.

**Table 2.** Results from the SVR kernel method investigation, including both vectorial and graph kernels. Included in the table are the values for the optimal hyperparameters for each SVR model (C, gamma and epsilon). The results of each polynomial degree of free (linear, quadratic and cubic) as well as the optimisation of n\_iter parameter for the WLS/WLOA were included to give total transparency in how each model was built. The models highlighted in green indicate the highest performing models which have plots of predicted vs experimental in figure 2

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Descriptor | Kernel | C | g | e | Degree | n\_iter | Training set | | Test set | | Validation set | |
| r2 | RMSE | r2 | RMSE | r2 | RMSE |
| ESOL | Linear | 2-5 | - | 1 | - | - | 0.56 | 1.55 | 0.55 | 1.61 | 0.63 | 1.24 |
| RBF | 2 | 2-1 | 1-1 | - | - | 0.70 | 1.27 | 0.66 | 1.38 | 0.72 | 1.07 |
| Poly | 2-4 | 2-3 | 1 | 1 | - | 0.56 | 1.55 | 0.55 | 1.61 | 0.63 | 1.24 |
| Poly | 2-4 | 2-1 | 1 | 2 | - | 0.14 | 2.16 | 0.17 | 2.16 | 0.14 | 1.89 |
| Poly | 20 | 2-5 | 1 | 3 | - | 0.19 | 2.09 | 0.21 | 2.12 | 0.20 | 1.83 |
| Sigmoid | 24 | 2-11 | 1 | - | - | 0.56 | 1.55 | 0.55 | 1.61 | 0.63 | 1.24 |
| MACCS | Linear | 2-5 | - | 1 | - | - | 0.61 | 1.45 | 0.58 | 1.55 | 0.49 | 1.3 |
| RBF | 22 | 2-7 | 1-1 | - | - | 0.88 | 0.80 | 0.72 | 1.25 | 0.67 | 1.03 |
| Poly | 2-5 | 2-1 | 1 | 1 | - | 0.61 | 1.46 | 0.58 | 1.55 | 0.49 | 1.29 |
| Poly | 2-3 | 2-4 | 1-1 | 2 | - | 0.87 | 0.85 | 0.66 | 1.39 | 0.50 | 1.29 |
| Poly | 2-5 | 2-5 | 1-1 | 3 | - | 0.87 | 0.83 | 0.69 | 1.33 | 0.57 | 1.18 |
| Sigmoid | 25 | 2-13 | 1 | - | - | 0.6 | 1.47 | 0.57 | 1.56 | 0.50 | 1.28 |
| ECFP | Linear | 2-5 | - | 1-1 | - | - | 0.71 | 1.24 | 0.51 | 1.67 | 0.35 | 1.46 |
| RBF | 2 | 2-9 | 1-2 | - | - | 0.85 | 0.87 | 0.36 | 1.91 | 0.23 | 1.59 |
| Poly | 2-5 | 2-7 | 1 | 1 | - | 0.63 | 1.41 | 0.54 | 1.62 | 0.53 | 1.24 |
| Poly | 25 | 2-6 | 1-1 | 2 | - | 0.99 | 0.22 | 0.43 | 1.79 | 0.18 | 1.64 |
| Poly | 2-5 | 2-6 | 1-1 | 3 | - | 0.97 | 0.42 | 0.4 | 1.85 | 0.45 | 1.34 |
| Sigmoid | 23 | 2-15 | 1 | - | - | 0.68 | 1.31 | 0.56 | 1.57 | 0.49 | 1.30 |
| Graphical | NH | 23 | 2-15 | 1-1 | - | - | 0.97 | 0.37 | 0.77 | 1.13 | 0.74 | 0.92 |
| RW | 25 | 2-15 | 1 |  | - | 0.00 | 2.39 | 0.0 | 2.39 | -0.08 | 1.89 |
| PG | 23 | 2-15 | 1 | - | - | 0.56 | 1.55 | 0.45 | 1.77 | 0.35 | 1.47 |
| ODD | 2 | 2-15 | 1-1 | - | - | 0.73 | 1.20 | 0.51 | 1.66 | 0.52 | 1.25 |
| PM | 22 | 2-15 | 1 | - | - | 0.71 | 1.25 | 0.64 | 1.43 | 0.5 | 1.28 |
| SP | 2-2 | 2-15 | 1 | - | - | 0.25 | 2.02 | 0.23 | 2.09 | 0.02 | 1.80 |
| GS | 21 | 2-15 | 1 | - | - | 0.06 | 2.25 | 0.02 | 2.36 | -0.01 | 1.82 |
| VH | 21 | 2-15 | 1 | - | - | 0.25 | 2.02 | 0.25 | 2.15 | 0.18 | 1.68 |
| WLS | 4 | 2-15 | 1 | - | 1 | 0.53 | 1.59 | 00.47 | 1.73 | 0.48 | 1.3 |
| WLS | 32 | 2-15 | 1 | - | 2 | 0.74 | 1.18 | 0.61 | 1.49 | 0.58 | 1.2 |
| WLS | 16 | 2-15 | 1 | - | 3 | 0.85 | 0.89 | 0.69 | 1.33 | 0.66 | 1.1 |
| WLS | 16 | 2-15 | 1 |  | 4 | 0.91 | 0.7 | 0.70 | 1.3 | 0.68 | 1.2 |
| WLS | 16 | 2-15 | 1 | - | 5 | 0.93 | 0.6 | 0.70 | 1.31 | 0.69 | 1.0 |
| WLS | 8 | 2-15 | 1 | - | 6 | 0.93 | 0.64 | 0.71 | 1.29 | 0.69 | 1.0 |
| WLS | 8 | 2-15 | 1 | - | 7 | 0.94 | 0.58 | 0.70 | 1.3 | 0.69 | 1.0 |
| WLS | 8 | 2-15 | 1 | - | 8 | 0.95 | 0.54 | 0.70 | 1.31 | 0.69 | 1.0 |
| WLS | 8 | 2-15 | 1 | - | 9 | 0.95 | 0.51 | 0.70 | 1.31 | 0.69 | 1.0 |
| WLS | 8 | 2-15 | 1 | - | 10 | 0.95 | 0.5 | 0.70 | 1.31 | 0.69 | 1.0 |
| WLS | 8 | 2-15 | 1 | - | 11 | 0.96 | 0.48 | 0.69 | 1.31 | 0.70 | 1.0 |
| WLS | 8 | 2-15 | 1 | - | 12 | 0.96 | 0.48 | 0.69 | 1.34 | 0.70 | 1.0 |
| WLS | 8 | 2-15 | 1 | - | 13 | 0.96 | 0.48 | 0.60 | 1.34 | 0.70 | 1.0 |
| WLS | 8 | 2-15 | 1 |  | 14 | 0.96 | 0.47 | 0.69 | 1.34 | 0.70 | 1.0 |
| WLS | 8 | 2-15 | 1 | - | 15 | 0.96 | 0.46 | 0.68 | 1.34 | 0.70 | 1.0 |
| WLOA | 2 | 2-15 | 1 | - | 1 | 0.74 | 1.18 | 0.71 | 1.29 | 0.63 | 1.1 |
| WLOA | 4 | 2-15 | 1-1 | - | 2 | 0.85 | 0.91 | 0.77 | 1.15 | 0.69 | 1 |
| WLOA | 4 | 2-15 | 1-1 | - | 3 | 0.9 | 0.72 | 0.78 | 1.11 | 0.74 | 0.93 |
| WLOA | 2 | 2-15 | 1-1 | - | 4 | 0.9 | 0.74 | 0.78 | 1.11 | 0.75 | 0.91 |
| WLOA | 2 | 2-15 | 1-1 | - | 5 | 0.92 | 0.64 | 0.78 | 1.12 | 0.75 | 0.91 |

**Figure 2.** Predicted vs experimental plots for best performing models for ESOL RBF, MACCS RBF, MACCS POLY(3), ECFP RBF, NH Graph, WLS Graph, WLOA Graph. Predicted vs experimental plots for the best performing models highlighted in table 2 to understand how well each model fits the data. Dark blue plot (left) indicates training set, light blue plot (centre) indicates test set and green plot (right) indicates validation set. Black fit line indicates what the theoretical fit should be if the model perfectly predicted. Red fit line is the actual fit observed for the model.

**Chart, scatter chart

Description automatically generated**Chart, scatter chart

Description automatically generated

**Chart, scatter chart

Description automatically generated**Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated**Chart, scatter chart

Description automatically generated**

**Figure 5.** Predicted (y-axis) vs experimental (x-axis) for the (A) MACCS RBF (B) NH (C) ESOL RBF (D) WLS models on the SC2 dataset. Blue fit indicates theoretical fit if the model perfectly predicted solubility. The red fit line indicates the actual plot of the models

(B)

(C)

(D)

**Chart, scatter chart

Description automatically generated**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

(A)