AI 4 Science Discovery Network+

Group: 1
Challenge: Task 1 - Predict solubility given a large set of calculated features
AI4SD ML Summer School Report
20-24th June 2022

Project Team: Jonathan Swain (University of Cambridge), Bradley Patrick (Nottingham Trent University), Andrea Frisco (University College London), Dan Criveanu (University of Nottingham)

Report Date: 28/06/2022

AI4SD-SummerSchool-Series:Report-1

Group: 1
Challenge: Task 1 - Predict solubility given a large set of calculated features
AI4SD-SummerSchool-Series:Report-1
Report Date: 28/06/2022
DOI: 10.5258/SOTON/AI3SD0244
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Project Details 1

2 Project Team 1

2.1 Project Student 1

2.2 Challenge Description 1

3 Lay Summary 1

4 Methodology 2

5 Results 2

6 Conclusions & Future Work 5

7 Outputs, Data & Software Links 5

8 Literature Survey 6

References 7

List of Figures 8

1 Project Details

Group Number	1
Challenge Name	Task 1 - Predict solubility given a large set of calculated features
Project Dates	20-24th June 2022

2 Project Team

2.1 Project Student

Name and Title	Jonathan Swain
Employer name / University Department Name	University of Cambridge
Work Email	jas272@cantab.ac.uk

Name and Title	Bradley Patrick
Employer name / University Department Name	Nottingham Trent University
Work Email	brad.patrick@ntu.ac.uk

Name and Title	Andrea Frisco
Employer name / University Department Name	University College London
Work Email	andrea.friso.21@ucl.ac.uk

Name and Title	Dan Criveanu
Employer name / University Department Name	University of Nottingham
Work Email	pcydc6@nottingham.ac.uk

2.2 Challenge Description

Task 1 - Predict solubility given a large set of calculated features.

3 Lay Summary

Solubility is one of the most important factors to consider during drug discovery. The ability of three linear regression models to predict solubility was assessed using the data set provided. The data set provided The models assessed used L2 and L1 regularization, and the third model was the Partial Least Squares (PLS) regression model. PLS regression was found to be the best

model out of the three models assessed. The results from using PLS regression on the test data set were further analysed and outliers were identified. The outliers were characterised, and it was found that the PLS regression model has issues predicting the solubility of highly aromatic compounds.

4 Methodology

Tools used:

- Overleaf / LaTeX
- Teams
- Github
- Jupyter Notebook
- Mendeley
- Powerpoint

Three models were used to predict solubility in terms of the logarithm of solubility (log S). The three models used utilised L1 and L2 regularisation, while the third model was the Partial Least Squares (PLS) regression model. For all three models, 30% of the Husskonen data set was used as the training set. The models were evaluated by calculating the root mean square error (RMSE) of the estimated log S values as compared to the actual log S values.

Outliers were defined as the values with an absolute residual greater than 1 on the estimated value-actual value plot........

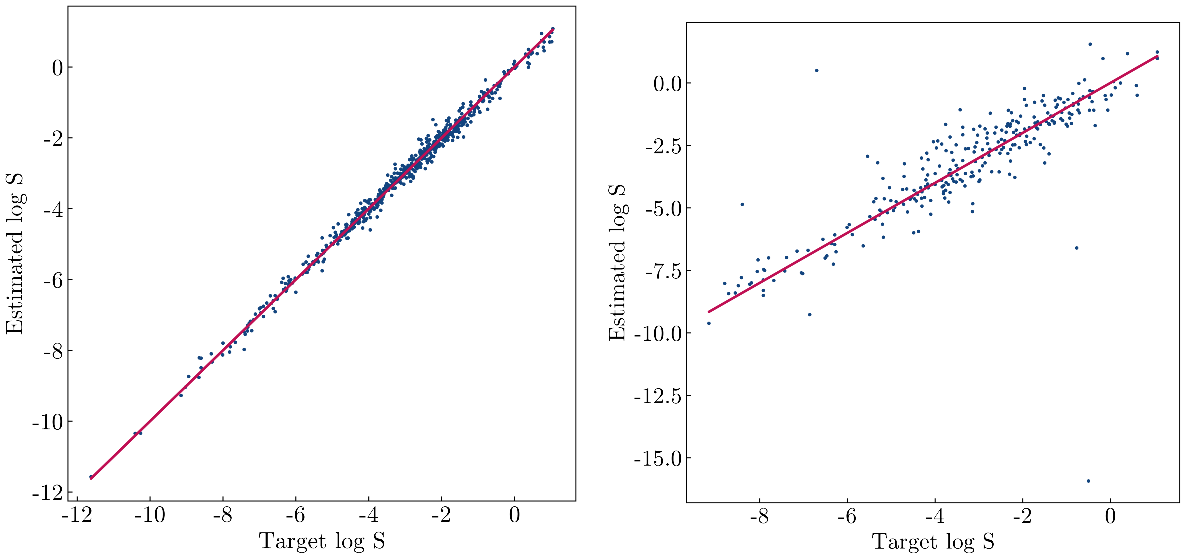
Missing information about clustering

5 Results

Linear regression with L2 regularization was used as the first model to predict solubility. The regularization parameter for L2 regularization, γ, was set to 2.3. On the training set, the RMSE was 0.17. On the test set, the RMSE was 1.32. Plots of the estimated values of log S against the actual values of log S on the training set and the test set can be seen in Figure 1a and 1b respectively. Based on the low RMSE on the training set, it is therefore clear that using L2 regularization has led to overfitting. Furthermore, given that solubility in the Husskonen data set is given in terms of the logarithm of solubility, an RMSE of 1.32 on the test set is not ideal. This poor performance could be due to some of the features in the data set being correlated.

In an attempt to avoid overfitting, linear regression with L1 regularization was subsequently employed. L1 regularization was chosen as it promotes sparsity, and thus acts as a form of feature selection. Feature selection is a method of avoiding overfitting. The regularization parameter for L1 regularization, α, was set to 2.3. On the training set, the RMSE was 0.87. On the test set, the RMSE was 0.90. Plots of the estimated values of log S against the actual values of log S on the training set and the test set can be seen in Figure 2a and 2b respectively. These results show that L1 regularization has addressed the problem of overfitting that L2 regularization displayed. Linear regression with L1 regularization also has a higher accuracy than L2 regularization, as shown by the dramatic decrease in RMSE from 1.32 for L2 and 0.90 for L1 regularization.

The final model used was the PLS regression model. PLS component analysis was done with respect to mean square error (MSE), and it was found that using 7 PLS components was optimal. Therefore, 7 PLS components were used for this model. On the training set, the RMSE was 0.53. On the test set, the RMSE was 0.73. As for the previous two linear regression



(a)

(b)

Figure 1: Estimated values vs. actual values of the logarithm of solubility (log S) as predicted by linear regression with L2 regularization. The regularization parameter, γ, was equal to 2.3. **(a)** Training set. **(b)** Test set. The red line is the ideal fit.

methods, plots of the estimated values of log S against the actual values of log S on the training set and the test set have been computed and can be seen in Figure 4a and 4b respectively. With an RMSE of 0.73, PLS regression is the best model considered in this project, and it was thus used for further analysis. The RMSEs of each of the three linear regression models on the training and test data sets are summarised in Table 1.

Data set	L2	L1	PLS
Training	0.17	0.87	0.53
Test	1.32	0.90	0.73

Table 1: The root mean square error (RMSE) of linear regression models with L2 and L1 regression and the Partial Least Squares (PLS) regression model on the training and test data sets.

The outliers were output as a list of SMILES strings, which were converted to RDKit objects. Within the outliers dataset there was one invalid SMILES, which was excluded leaving 41 structures. A quick visual examination of the structures showed no apparent similarities between the structures, so a clustering analysis was completed [1].

Using RDKit, RDK5 molecular fingerprints were created for each compound, each containing 2048 possible fragments. The similarity between two fingerprints was calculated using the Tanimoto Coefficient [2]. Completing this for all pairs of fingerprints allowed creation of a similarity matrix between all molecular fingerprint pairs. Subtraction of each of these values from 1 created a distance matrix between all molecular fingerprint pairs.

Butina clustering was completed on the distance matrix using RDKit [3]. Experimenting with varying cut-offs from 0.0 to 1.0 in steps of 0.2 found that a cut-off of at least 0.8 was required for clusters of at least five compounds. Visual examination of the two largest clusters showed all compounds contained aromatic systems, as seen in Figure 6. Further work is required to determine how to improve the model to account for aromatic compounds. The initial step

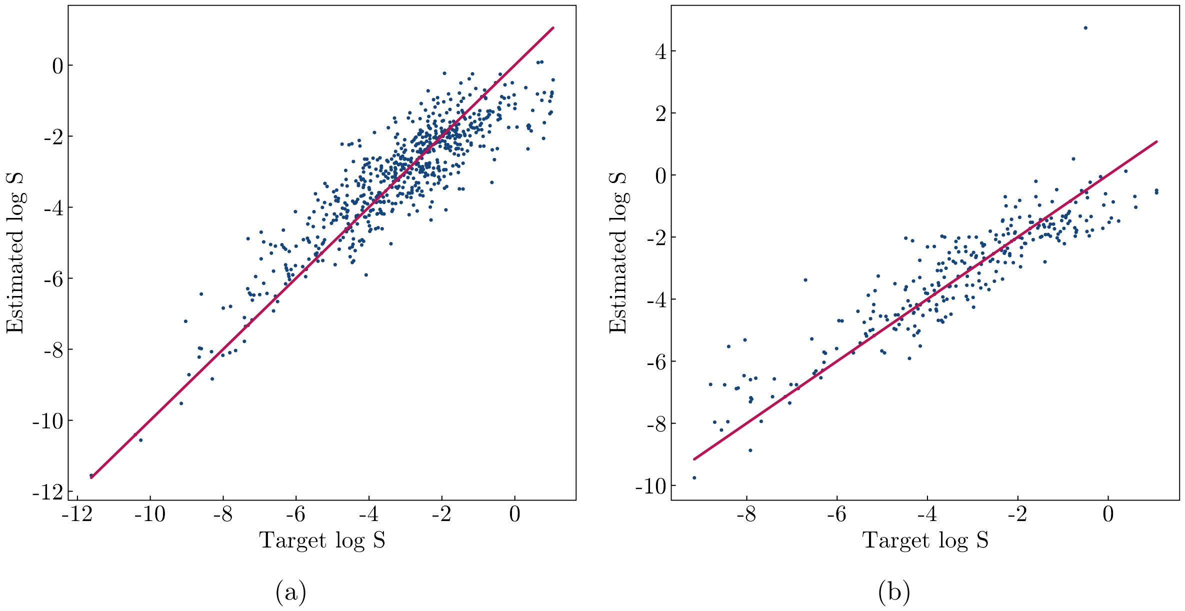


Figure 2: Estimated values vs. actual values of the logarithm of solubility (log S) as predicted by linear regression with L1 regularization. The regularization parameter, α, was equal to 2.3. **(a)** Training set. **(b)** Test set. The red line is the ideal fit.

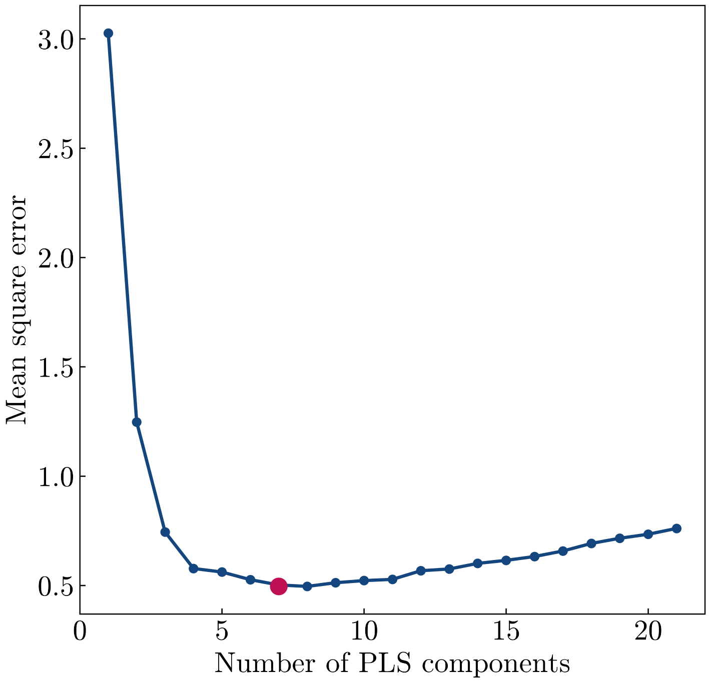


Figure 3: A plot showing how mean square error (MSE) varies with the number of Partial Least Squares (PLS) components. The red circle highlights the optimal number of PLS components, i.e. the number with the lowest MSE, 7.

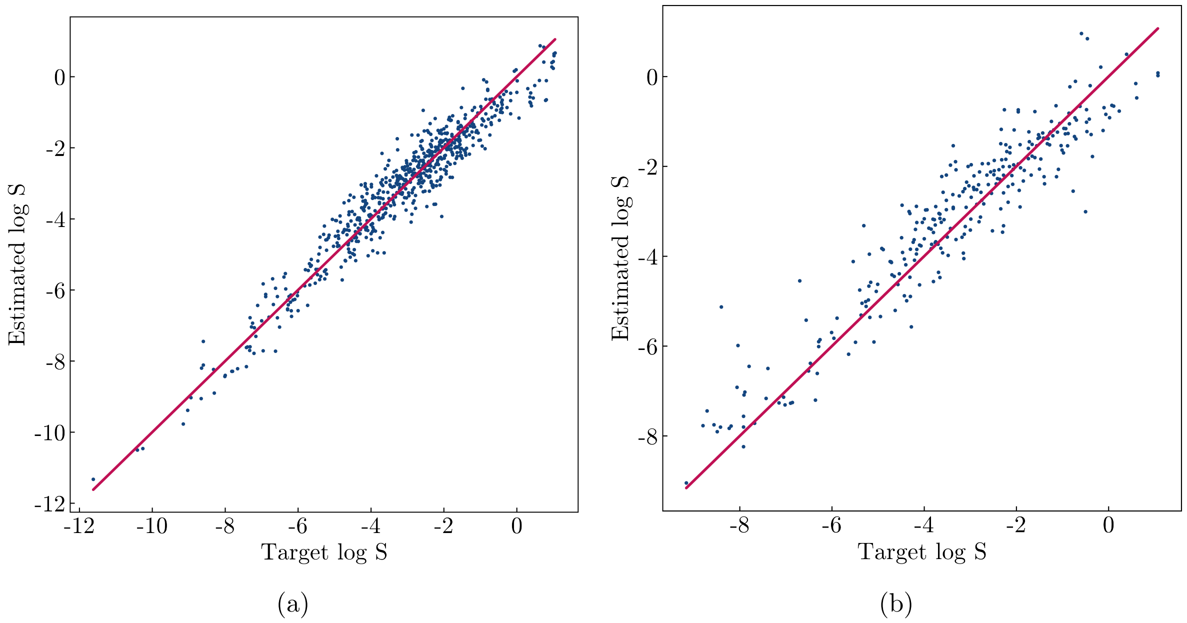


Figure 4: Estimated values vs. actual values of the logarithm of solubility (log S) as predicted by Partial Least Squares (PLS) regression. Seven PLS components were used. **(a)** Training set. **(b)** Test set. The red line is the ideal fit.

will be to determine if the model is consistently over-, or underestimating the solubility of these compounds.

6 Conclusions & Future Work

Guidance (Delete upon completion): Please provide the conclusions for the work carried out in this challenge, and include any relevant comments and ideas on how this work could be continued in the future.

The dataset used consisted of roughly 935 possible molecules

The success can be found based on the accuracy...

The most important features are...

The Husskonen dataset has approx. 935 molecules in comparison, which is significantly smaller compared to the 17,000. Due to the limited size, Deep Learning would not be possible.

In this report, three linear regression models were assessed on their suitability to predict solubility: L1 and L2 regularization, and PLS regression. It was found that PLS regression had the lowest RMSE on the test set out of the three models, and was thus the best model considered in this report. Using the results of using PLS regression on the test set, outliers were identified and characterised. It was found that the most common type of molecule that PLS regression struggles to correctly predict the solubility of were highly aromatic molecules.

Regarding future work, other models, such as the state-of-the-art models based on deep learning, could be tried on the Husskonen data set.

7 Outputs, Data & Software Links

Guidance (Delete upon completion): If you have produced any outputs (including poster presentations, GitHub repositories of data or software produced etc) please include details and links to these.

GitHub repository: `https://github.com/brad-pat/AI_G1`

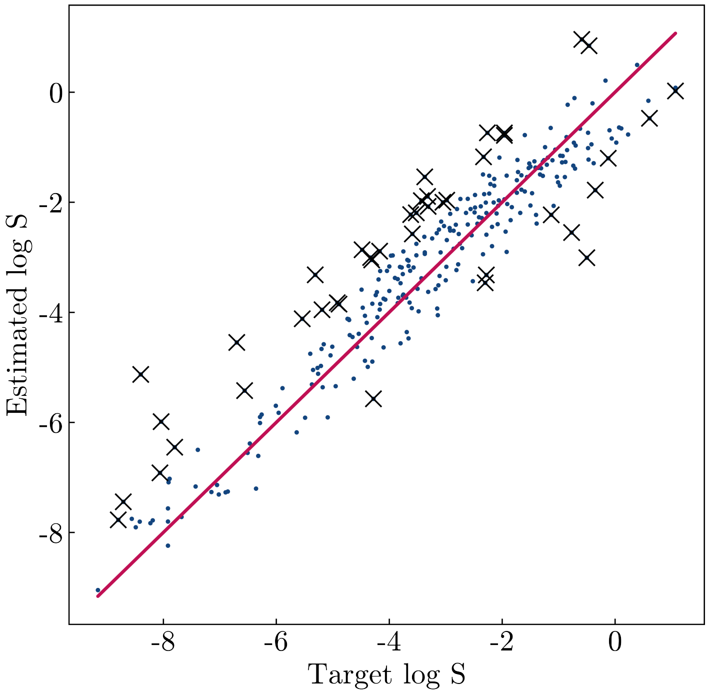


Figure 5: An estimated vs. target value plot of the logarithm of solubility (log S), highlighting the outliers (black crosses) identified using Partial Least Squares (PLS) regression on the test data set. The red line shows the ideal fit.

Dropbox: https://myntuac-my.sharepoint.com/:f:/g/personal/brad_patrick_ntu_ac_uk/EioMQ7xfsc1FhqQ5CagEyDYBDxcvB_XR22r-Ee0D4pskRg?e=sNGUhS

8 Literature Survey

For this section, some research was conducted to check how molecule solubility prediction has been done before. Solubility prediction has been an area of research that has been done for a while. There are several studies on this with a range of different data sets with different sizes. Based on the literature, however, Deep Learning seems to be one of the more common approaches tried in molecule prediction.

In [4], they predicted Aqueous solubility using several different Deep Learning (DL) models. These DL models were all performed on a molecule data set of over 17,000, which they claim is the most diverse collection of organic modules to date. In [5], they also had a relatively large data set with just under 10,000 molecules. They state that this is a limited data set and it did have an effect on the overall training for their models. Both of these studies attempt to implement a DL approach, but even with the large data set both studies have, it has been shown that DL requires a much larger data set to be able to gauge more meaningful predictability.

Outside the realms of DL, in [6], the authors created their own Machine Learning tool and evaluated it against a range of different data sets. They concluded that their tool outperforms linear Machine Learning approaches and matches other Machine Learning methods. The authors also state their concern that limited data sets not providing the greatest of results.

The Husskonen data set will be used for this experiment and has approximately 935 molecules. By comparison, it is significantly smaller compared to the 10,000 or 17,000 mentioned in the above studies. Due to this, DL would likely be an unsuccessful route due to the limited size of the data set that the experiment would be conducted with.

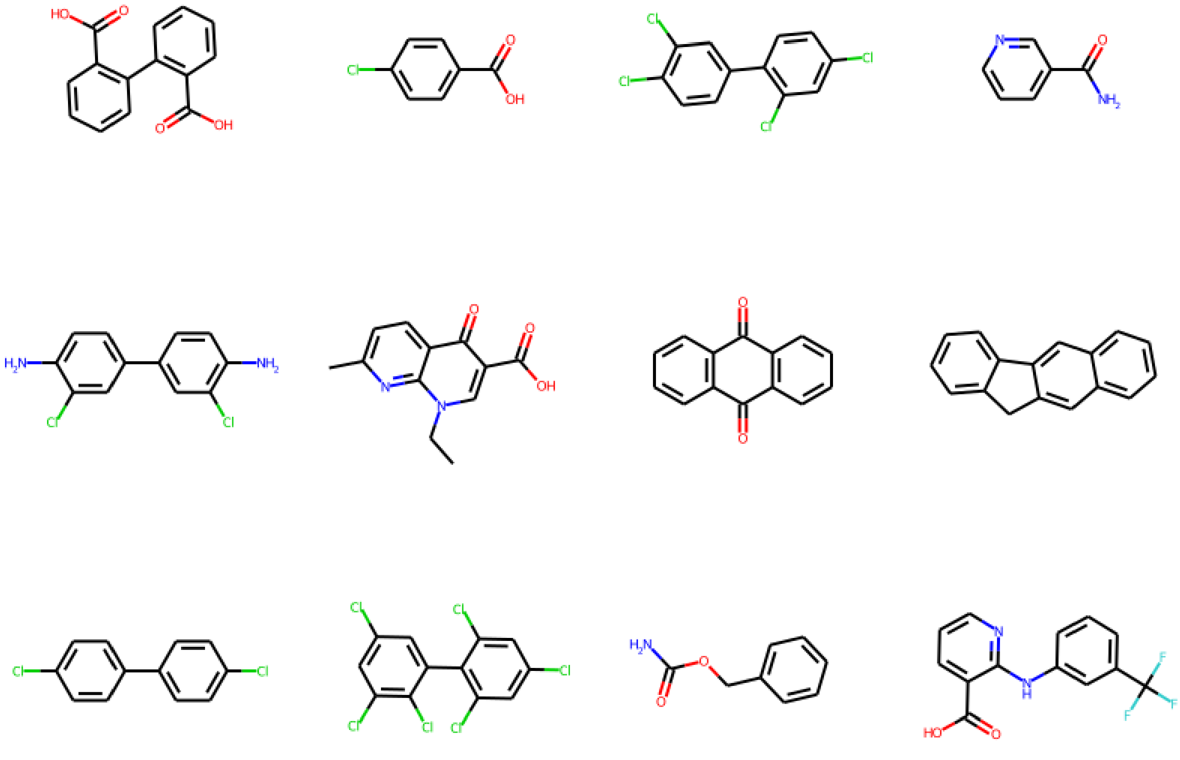


Figure 6: Chemical structures of the compounds found in the largest cluster of outliers, each containing at least one aromatic group.

References

[1] Spriewald G, Caswara C, Rodríguez-Guerra J. T005 · Compound clustering; 2022.

[2] Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? Journal of Cheminformatics. 2015 6;7(1).

[3] Butina D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. Journal of Chemical Information and Computer Sciences. 1999 6;39(4):747-50. Available from: https://doi.org/10.1021/ci9803381.

[4] Panapitiya G, Girard M, Hollas A, Murugesan V, Wang W, Saldanha E. Predicting Aqueous Solubility of Organic Molecules Using Deep Learning Models with Varied Molecular Representations. 2021 5. Available from: http://arxiv.org/abs/2105.12638.

[5] Cui Q, Lu S, Ni B, Zeng X, Tan Y, Chen YD, et al. Improved Prediction of Aqueous Solubility of Novel Compounds by Going Deeper With Deep Learning. Frontiers in Oncology. 2020 2;10.

[6] Francoeur PG, Koes DR. SolTranNet-A Machine Learning Tool for Fast Aqueous Solubility Prediction. Journal of Chemical Information and Modeling. 2021 6;61(6):2530-6.

List of Figures

1 Estimated values vs. actual values of the logarithm of solubility (log S) as predicted by linear regression with L2 regularization. The regularization parameter, γ, was equal to 2.3. **(a)** Training set. **(b)** Test set. The red line is the ideal fit. . 3

2 Estimated values vs. actual values of the logarithm of solubility (log S) as predicted by linear regression with L1 regularization. The regularization parameter, α, was equal to 2.3. **(a)** Training set. **(b)** Test set. The red line is the ideal fit. . 4

3 A plot showing how mean square error (MSE) varies with the number of Partial Least Squares (PLS) components. The red circle highlights the optimal number of PLS components, i.e. the number with the lowest MSE, 7. 4

4 Estimated values vs. actual values of the logarithm of solubility (log S) as predicted by Partial Least Squares (PLS) regression. Seven PLS components were used. **(a)** Training set. **(b)** Test set. The red line is the ideal fit. 5

5 An estimated vs. target value plot of the logarithm of solubility (log S), highlighting the outliers (black crosses) identified using Partial Least Squares (PLS) regression on the test data set. The red line shows the ideal fit. 6

6 Chemical structures of the compounds found in the largest cluster of outliers, each containing at least one aromatic group. 7