Full Paper

www.molinf.com

molecular informatics

DOI: 10.1002/minf.201500052

Predicting Melting Points of Organic Molecules: Applications to Aqueous Solubility Prediction Using the General Solubility Equation

J. L. McDonagh,[a, b] T. van Mourik,[a] and J. B. O. Mitchell*[a]

Abstract: In this work we make predictions of several important molecular properties of academic and industrial importance to seek answers to two questions:
1) Can we apply efficient machine learning techniques, using inexpensive descriptors, to predict melting points to a reasonable level of accuracy?
2) Can values of this level of accuracy be usefully applied to predicting aqueous solubility?
We present predictions of melting points made by several novel machine learning models, previously applied to solubility prediction. Additionally, we make predictions of solubility via the General Solubility Equation (GSE) and monitor the impact of varying the logP prediction model (AlogP and XlogP) on the GSE. We note that the machine learning models presented, using a modest number of 2D descriptors, can make melting point predictions in line with the current state of the art prediction methods (RMSE $\geq$ 40 °C). We also find that predicted melting points, with an RMSE of tens of degrees Celsius, can be usefully applied to the GSE to yield accurate solubility predictions ($\log_{10}$S RMSE $<$ 1) over a small dataset of drug-like molecules.

Keywords: Machine learning · Melting points · Pharmaceuticals · QSPR · Solubility

1 Introduction

We investigate the prediction of two important molecular properties, melting point and aqueous solubility, in this work. Melting points can provide information on the stability of a crystal's structure and provide insights into polymorphism, using experimental techniques such as differential scanning calorimetry. The enantiotropy or monotropy of a polymorphic system can be investigated by cycles of melting and recrystallization.[1] This situation can be more complicated when samples are impure or solvated structures arise from the recrystallization process. This makes melting points important pieces of industrial knowledge. Melting points can also offer an insight into other chemical properties. One such property is aqueous solubility, via the well-established General Solubility Equation (GSE; Equation 1),[2] which links the melting point to solubility with reference to a thermodynamic cycle via a pure melt (Figure 1).[3] The GSE approximates this thermodynamic cycle using the melting point to approximate ΔG_{fusion} and logP to approximate $\Delta G_{transfer}$. This relationship has seen wide usage.[2b,4]

$$\log_{10} S = 0.08 - \log_{10} P - 0.01 \times (MP - 25)$$

$$\log_{10} S = 0.05 - \log_{10} P - 0.01 \times (MP - 25)$$

Equation 1. Top: original GSE from Yalkowsky and Valvani,[2c] bottom: revised GSE by Jain and Yalkowsky.[2a] $\log_{10}$S is the logarithm to the base ten of the aqueous solubility (S; units referred to mol/L), $\log_{10}$P is the base ten logarithm of the n-octanol/water partition coefficient, and MP (°C) is the melting point. The two versions of the GSE are used without modification throughout this work.

The GSE was proposed as a method to accurately predict solubility using only two pieces of empirical data; the first is the melting point, the second logP. LogP can be reasonably predicted by atom contribution models such as AlogP[5] and XlogP[6] or group contribution models such as ClogP.[7] Melting points, however, still elude us as predictable quantities. For this reason, a good prediction of a crystal's melting point could in principle provide a more accurate prediction of a molecule's solubility, noting that within the framework of the GSE the logP term will always be the dominant contribution.

A variety of methods have been trialled previously to predict melting points; these have included Quantitative Structure-Property Relationships (QSPR),[8] several attempts

[a] J. L. McDonagh, T. van Mourik, J. B. O. Mitchell
School of Chemistry, University of St Andrews, North Haugh, St Andrews, Fife, Scotland, United Kingdom, KY16 9ST
*e-mail: jbom@st-andrews.ac.uk

[b] J. L. McDonagh
Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK

Supporting information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201500052.

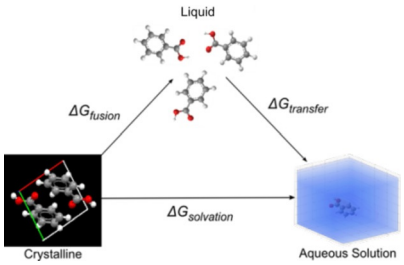

Figure 1. A thermodynamic cycle for the prediction of aqueous solubility via a super cooled liquid state.

of this kind have been made. Some notable examples are those of Bergstrom *et al.* who predicted 92 drug molecules' melting points using the Partial Least Squares (PLS) machine learning method, achieving an RMSE of 49.8 °C.[8b] Hughes *et al.* predicted the melting points of 287 drug-like molecules with a range of machine learning methods, achieving their best prediction with the support vector machine (SVM) model, RMSE 52.8 °C.[8a] Nigsch *et al.* predicted the melting points of 80 drug molecules with an RMSE of 46.3 °C.[8c] More recently, Tetko *et al.* have produced a method for predicting drug molecule melting points. They analysed 47,000 compounds and found that >90% of their melting points lay in the interval of 50–250 °C. The model aims to make accurate predictions in this interval. An associative neural network was used to build models validated using a 5-fold cross validation. A consensus model was then finally created by averaging over the results. This gave a final predictive accuracy of an average RMSE of 33 °C across the temperature interval of 50–250 °C.[9] A final recent QSAR study from Kew *et al.*[10] tests the predictive accuracy of a large range of machine learning methods using the dataset of Hughes *et al.*[8a] The study concludes with the best prediction method being a Greedy ensemble (an ensemble prediction based on the predictions made from several other machine learning models) using principal components analysis to pre-screen the input data. This method made a prediction with an RMSE of 52.9 °C over 287 data points and a temperature span of 350 °C.[10]

Recently, Preiss *et al.*[11] obtained results from fitted equations generated using physically motivated calculations. This work achieved an average error in prediction of 33.5 °C over 520 chemically diverse 1:1 organic salts.

To date, an accuracy of tens of degrees Celsius over reasonably sized datasets represents the state of the art. In this work we pose two questions:

i) Can we apply efficient machine learning techniques, using inexpensive descriptors, to predict melting points to a reasonable level of accuracy?
ii) Can values of this level of accuracy be usefully applied to predicting aqueous solubility?

We investigate the use of three machine learning methods to predict 1100 melting points. The data from these melting points comes from the Alfa Aesar open melting point dataset.[12] We proceed from these melting point predictions to predict the aqueous solubility, using the GSE, of a small subset of this melting point dataset which overlaps with our previously created Drug-Like Solubility-100 (DLS-100) dataset.[13]

2 Experimental

2.1 Melting point and logP data

Several groups have created open source melting point datasets which are available online (http://lxsrv7.oru.edu/~alang/meltingpoints/download.php). These data sets come from a variety of sources and publications over the years. The largest dataset from a single source presented here comes from Alfa Aesar. This data set, containing over 8000 molecules, is conveniently available with chemical names, SMILES strings, CSID's, weblinks and CAS numbers.[12]

In this work we take a subset totalling 1100 molecules from the original Alfa Aesar dataset. The first 1000 molecules in the dataset were taken followed by 100 chosen at random from the remaining dataset. This dataset was provided to several machine learning models for training and prediction of melting points. Each molecule in the Alfa Aesar dataset is given just a single value for their respective melting points. We will refer to this dataset as MP1100. MP1100 is presented with molecular name, logP predicted from the AlogP algorithm, and melting point in the electronic supporting information.

AlogP and ClogP are both widely used algorithms which predict the n-octanol/water partition coefficient. AlogP was selected in this work. This selection was based upon AlogP having previously been demonstrated to make predictions of equivalent accuracy to the ClogP algorithm, for molecules of the order of 21–45 atoms in size. Additionally, the AlogP algorithm demonstrated improved predictions compared to the ClogP algorithm for molecules consisting of >45 atoms. 67 molecules in MP1100 have an atom count of ≥45 atoms and 597 other molecules consist of between 21 and 45 atoms. We therefore considered the AlogP method the better choice.[14]

MP1100 is a diverse dataset, comprised of molecules covering a wide range of chemical space. MP1100 contains only organic molecules, several of which show structural isomerisation. All melting points and SMILES are used without modification directly from the Alfa Aesar dataset. The MP1100 melting points are approximately normally distributed. The melting point temperature range of MP1100 is between −142 °C to 375 °C, i.e. a total span of 517 °C (Figure 2).

2.2 Machine Learning

We use three machine learning models: Random Forest (RF),[15] Support Vector Machines (SVM)[16] and Partial Least

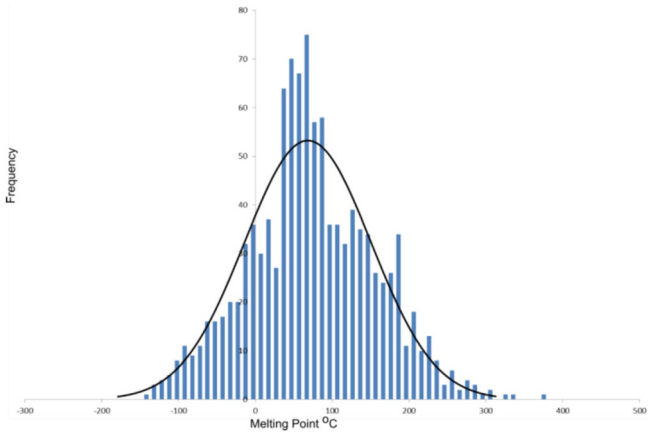

Figure 2. Distribution of the MP1100 melting point temperatures.

Squares (PLS).[17] These methods have been explained previously elsewhere[18] and so we restrict ourselves to a brief conceptual explanation of these methods.

RF is an ensemble learning method, which creates a forest of decision trees via recursive partitioning. This scheme aims to place data points with similar experimental values in the same final node (leaf node) of the tree. These decision trees are built using a random subset of the descriptors which are provided to the splitting algorithm, and chosen afresh for each node of the tree. This algorithm selects the optimal splitting based on some pre-defined criteria. The splitting criterion in this case was to minimise the RMSE. The predicted value associated with this leaf is given by the average of all training data which fall into the same leaf node.

SVM projects the data into a higher order feature space using a kernel function (a radial basis kernel in this case). The method attempts to fit a regression in the feature space by defining a hyper-plane which maximally explains the data. This is done whilst minimising the error margins of the regression to avoid overfitting.

PLS looks to generate a set of latent variables (linear combination of descriptors). This is done in such a way as to maximally explain the covariance between the dependent and independent variables. As the latent variables are linear combinations of descriptors there are always fewer latent variables than descriptors; this helps to avoid overfitting.

We used SMILES strings to calculate 2D descriptors from the open source Java library the Chemistry Development Kit (CDK).[19] A list of all available descriptors in the CDK and definitions can be found at the website listed in reference.[20] 101 descriptors in total provided information to the models. As this number of descriptors is considerably smaller than the number of data points, there should be little concern of overfitting. Auto-scaling was applied as an additional precaution. This scaling centres each descriptor on a mean value with a normalised standard deviation of one. This method of scaling was selected as the models we had previously created for solubility prediction had performed well using auto-scaling.[13]

The machine learning was carried out using R and a double tenfold cross validation methodology. We have previously applied a similar scheme, with some success, to solubility prediction.[13] Each of the machine learning methods have a number of internal parameters which require optimisation. In addition we must produce an unbiased training and test set separation and run this over three machine learning models. This can be achieved efficiently using a double ten-fold cross validation approach. This provides an internal ten-fold cross validation, in which the machine learning parameters are optimised and a second external ten-fold cross validation in which an unbiased split of the data is made into a test and training set. A scheme representing the key steps is given below (Figure 3). A more detailed scheme is provided in the electronic supporting information (Figure S1).

This process proceeds as follows for MP1100. For 1100 molecules the training and test set split is made in an unbiased way with a random selection of 10% of the data to be the test set. This is repeated ten times, hence 10-fold cross validation. The remaining 90% is used as training data. These training data are then split again. This time 10% (9% of the original data) are taken randomly to test the parameters being trialled in the machine learning models. The remaining 90% (81% of the original data) is used to train the machine learning models built using the test parameters. Optimised parameters are considered to be those minimising the RMSE of the internal 10-fold cross validation's test set. The scripts used to run such machine learning are provided freely for download at the following

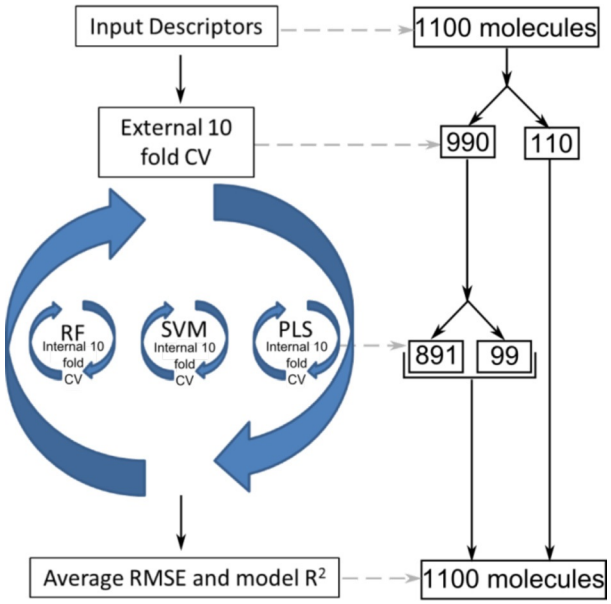

Figure 3. Double tenfold cross validation, internal 10-fold validation to optimise parameters and the external tenfold cross validation for model training and validation.

website (http://chemistry.st-andrews.ac.uk/staff/jbom/group/Informatics-Solubility.html).

2.2 Statistical Analysis

Throughout this work the following statistics are applied to analyse the resultant models. The first is the coefficient of determination (R^2), second the Root Mean Square Error (RMSE), third the standard deviation (σ), fourth the bias and finally the Average Absolute Error (AAE). The mathematical definitions of these statistical measures are:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_{exp}^{i} - y_{pred}^{i}\right)^2}{\sum_{i=1}^{n}\left(y_{pred}^{i} - \bar{y}\right)^2}$$

Equation 2. y_{exp} is the experimental value, y_{pred} is the predicted value and $\bar{y}$ is the mean

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{exp}^{i} - \bar{y}\right)^2}{N}}$$

Equation 3. y_{exp} is the experimental value, y_{pred} is the predicted value and N is the number of data points.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}\left(y_{exp-pred}^{i} - \bar{y}\right)^2}{N-1}}$$

Equation 4. $y_{exp-pred}$ is the difference between the experimental and predicted value, N is the number of data points and $\bar{y}$ is the mean experimental value.

$$Bias = \frac{\sum_{i=1}^{n}\left(y_{exp}^{i} - y_{pred}^{i}\right)}{N}$$

Equation 5. y_{exp} is the experimental value, y_{pred} is the predicted value and N is the number of data points.

$$AAE = \frac{\sum_{i=1}^{n}\left|y_{exp}^{i} - y_{pred}^{i}\right|}{N}$$

Equation 6. y_{exp} is the experimental value, y_{pred} is the predicted value and N is the number of data points. Note the numerator is sum of the absolute value of each difference.

We use these statistics to quantitatively analyse and compare our models. The R^2 value measures how well the model is fitting the data. RMSE measures the differences between the predicted values and the actual values, giving a model average error. RMSE can be thoughts of as an overall error; we can deepen our understanding of the error by consideration of the bias and σ. Here we consider the bias to be an estimate of the systematic error and σ an estimate of the random error of model. The AAE is used to quantify the average error over a set of predictions.

3 Results and Discussion

3.1 Melting Point Prediction

Given that previous work has shown logP to be an important descriptor for solubility prediction,[13,21] not least suggested in the GSE where logP is a core parameter, melting point predictions were run twice. In the first run the descriptor set contained a logP descriptor, whereas the second run did not contain a logP descriptor. This was to assess the numerical importance of a logP descriptor to the final models. The melting point is a property of the solid and molten liquid state. The a priori assumption was therefore that the log P descriptor, which describes the solvated phase, would have a negligible impact on melting point prediction. The results (shown in the electronic supporting information Figures S2–S5) in principle support this a priori hypothesis, although a small improvement in the melting point prediction is seen when a logP descriptor is included. In the rest of this article predictions are therefore made with the logP descriptor included.

3.2 MP1100 Melting Point Predictions

Figures 4–6 and Table 1 show the predictions of the MP1100 melting points from RF, SVM and PLS.

Table 1 shows that each of the machine learning methods achieves a good R^2 correlation coefficient indicating

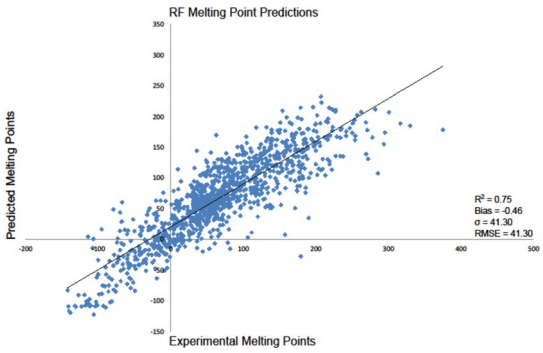

Figure 4. RF predictions of MP1100's melting points.

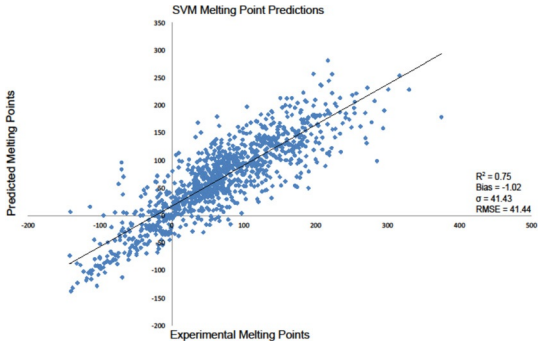

Figure 5. SVM predictions of MP1100's melting points.

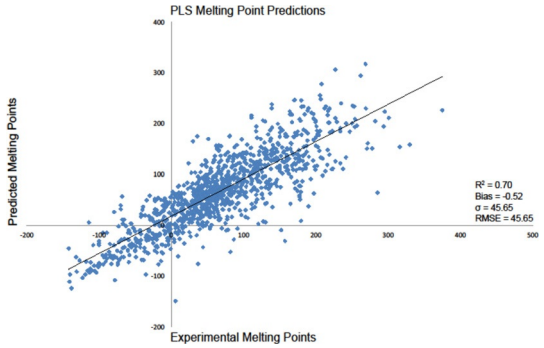

Figure 6. PLS predictions of MP1100's melting points.

Table 1. Prediction RMSE's and correlation coefficients for MP1100 melting points by RF, SVM and PLS.

Method	R^2	RMSE (°C)
RF	0.75	41.30
SVM	0.75	41.44
PLS	0.70	45.65

a good correlation between the predicted values and the experimental results. The bias, taken here to be a measure of systematic error, is also low in all cases. The experimental data has a standard deviation of 82.40 °C; all three models have an RMSE some way below this. We therefore suggest the models are making a useful prediction, i.e. better than the null model of the mean value.

However, the RMSE of the predictions is above 40 °C for all three models. This means that, at best, the results represent imprecise predictions, although these are of a similar accuracy to those discussed in the introduction. The average absolute predictive errors are: 34.4 °C from PLS, 30.7 °C from RF and 30.6 °C from SVM. Given that these predictions are over a large temperature range (517 °C), on average the predictive accuracy achieved is reasonably promising for a simple model using only a modest number of 2D descriptors. However, at the upper end of the predictive inaccuracy, one can find errors up to 220.8 °C for PLS, 207.5 °C for RF and 195.9 °C for SVM. Clearly, such inaccuracies in prediction mean those values are not of quantitative use, although they could provide some level of qualitative value. Overall the predictions presented here are of similar accuracy as existing methodologies using a variety of QSPR/QSAR approaches and datasets containing similar molecules.[8,22]

We analysed the results obtained with the RF model to determine the most important descriptors. Whilst these results should not be over analysed for strict chemical meaning, one may find useful information in such analyses. We find that the most important descriptor for the RF model is the topological polar surface area. Second is the Zagreb index, third and fourth are respectively weighted paths of length 3 and 4 and fifth the hydrogen bond donor count. The first descriptor describes the polarity of a molecule's exposed surface. This is the area best accessible for direct interaction. The second, third and fourth most important descriptors provide information on the molecule's extent in terms of complexity and branching. The fifth most important descriptor, the hydrogen bond donor count, can provide some information, although very limited, about significant interactions within the solid state which are not necessarily provided by the four previous descriptors. Interestingly, the hydrogen bond acceptor count is ranked much lower down, outside of the top ten. The logP descriptor is also found outside of the top ten. As we stated above, a logP descriptor is unlikely to provide significant information when trying to predict the melting point. The ranked top ten most important descriptors are shown in the electronic supporting information Table S2.

At this point we can respond to our first question ('can we apply efficient machine learning techniques, using inexpensive descriptors, to predict melting points to a reasonable level of accuracy?'), with the answer "nearly". Taking efficient and fairly simple machine learning models with a modest number of 2D cheminformatics descriptors; it is possible to predict melting points to a similar level of accuracy as some more complex methods. This can be considered a reasonable level of accuracy depending on one's end goal – what one wants to use the melting points for. While the predicted melting points are not quantitatively accurate, we show below that such predictions are suitably accurate for quantitative prediction of solubility via the GSE.

3.3 Solubility Prediction Using the GSE and Predicted Melting Points

Following from these predictions we considered the overlap between MP1100 and the existing solubility datasets, DLS-25 and DLS-100, from our own previous work.[13,23] The values in DLS-25 and DLS-100 were carefully curated from the literature using the criteria stated in the following references.[13,23] Where possible, these data come from the CheqSol method,[24] which has been shown to be robust, reliable and reproducible in determining a molecule's solubility. An overlap of 30 molecules was found and will be referred to as SOL-30.

Reliable solubility values are often hard to find and verify. We therefore chose here to use a small but reliable dataset as opposed to a larger but potentially very noisy dataset from a wider selection of literature sources. A reliable dataset should show more clearly the trends in the model which can be hidden in noise for a less reliable but larger dataset.

Predictions of the aqueous solubility (logS) were made for SOL-30 using the GSE. The GSE has two forms: a revised version from Jain and Yalkowsky[2a] and the original form, from Yalkowsky and Valvani.[2c] We present here the results of both forms for the same dataset. The results are sum-

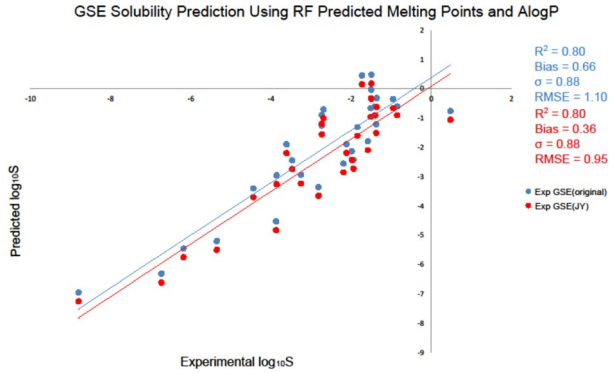

Figure 7. A prediction of solubility for the SOL-30 molecules using the General Solubility Equation with predicted melting points from RF and predicted logP from AlogP. Bias, standard deviation (SD) and RMSE are quoted in logS units.

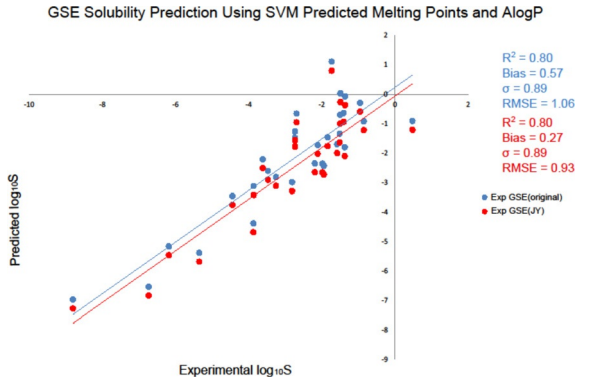

Figure 8. A prediction of solubility for the SOL-30 molecules using the General Solubility Equation with predicted melting points from SVM and predicted logP from AlogP. Bias, standard deviation (SD) and RMSE are quoted in logS units.

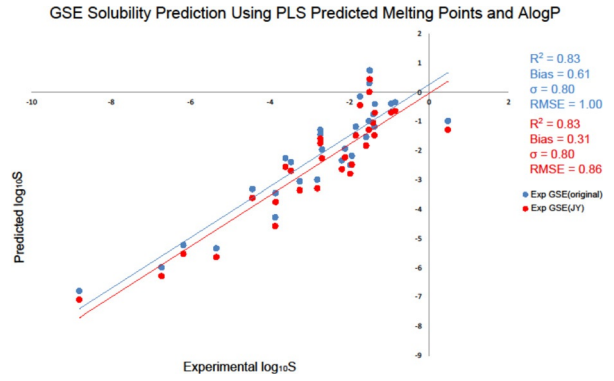

Figure 9. A prediction of solubility for the SOL-30 molecules using the General Solubility Equation with predicted melting points from PLS and predicted logP from AlogP. Bias, standard deviation (SD) and RMSE are quoted in logS units.

marised in Figures 7, 8 and 9. The results presented in Figures 7, 8 and 9 were obtained using the AlogP algorithm to calculate the logP.

The standard deviation of the experimental data is calculated to be 1.95 logS units. Table 2 summarises the numerical accuracy of the methods. We find that the revision of the GSE has a noteworthy effect on the bias, producing a substantial reduction in the systematic error on comparison with the GSE's original formulation. For solubility prediction, a useable level of accuracy is generally taken to be that of approximately 1 logS unit (5.7 kJ/mol) RMSE. This criterion is beaten for all predictions with the revised GSE and met or marginally missed by the original GSE. This is a very promising result for a simple model which uses melting points with an RMSE of the order of approximately 40 °C. The model utilising PLS predicted melting points performs very well. It shows low systematic and random errors. It is also the only method of the three in which the 1 logS unit accuracy target is met with the original form of the GSE.

Figure 10 shows solubilities predicted using the experimental melting points in place of the predicted melting points. As may be expected, there is an improvement in the RMSE scores and a diminishing of the random and systematic errors. Both forms of the GSE now meet the 1 logS accuracy criterion. This provides intuitive confirmation that the method has systematic improvement when given data of a higher accuracy.

The above finding is in agreement with some previous work.[8a] However, more recent work by Palmer *et al.*[25] has suggested that the quality of training data used in QSAR models is not the limiting factor in their accuracy. Palmer *et al.* find a small reduction in accuracy when reportedly more accurate experimental solubility data are provided to train a set of QSAR models, one of which is a multiple linear regression (MLR) model, based on MP and logP, inspired by the GSE. This is an important finding. Palmer *et al.* train two GSE inspired models using the same experimental MP and logP values against two different logS datasets, one reportedly more accurate than the other. In the present article, we compare the result of predicted and experimental data input into a GSE model. One may expect

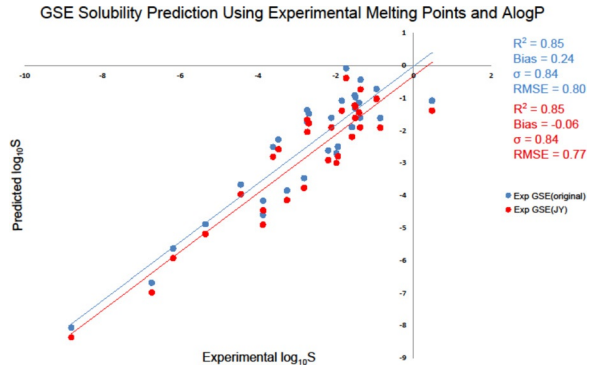

Figure 10. A prediction of solubility for the SOL-30 molecules using the General Solubility Equation with experimental melting points and predicted logP from AlogP.

Table 2. All data relating to predictions of melting point, AlogP and logS using the revised version of the GSE.

Molecule	Exp MP (°C)	Exp (logS)	AlogP	GSE_{JY} (Exp logS)	PLS MP (°C)	GSE_{JY} (PLS logS)	RF MP (°C)	GSE_{JY} (RF logS)	SVM MP (°C)	GSE_{JY} (SVM logS)
1,3,5-trichlorobenzene	63.5	−4.44	4.08	−3.97	28.3	−3.61	36.96	−3.7	42.39	−3.75
1−Naphthol	96	−1.98	2.79	−3	75.3	−2.79	38.66	−2.43	61.8	−2.66
4−Aminobenzoic acid	187.5	−1.37	0.78	−1.91	144.6	−1.48	147.66	−1.51	207.01	−2.1
5,5−Diphenylhydantoin	295.5	−3.86	2.26	−4.47	223.95	−3.75	174.2	−3.25	191.13	−3.42
Acetanilide	114.5	−1.4	1.05	−1.45	75.31	−1.05	59.92	−0.9	63.45	−0.93
Adenosine	235	−1.73	−1.21	−0.39	240.39	−0.44	181.05	0.15	115.41	0.81
Antipyrine	112.5	0.48	1.01	−1.39	102.4	−1.28	79.32	−1.05	94.7	−1.21
Benzamide	127	−0.95	0.51	−1.03	93.34	−0.69	89.24	−0.65	82.5	−0.58
Benzoic acid	122.5	−1.58	1.72	−2.2	85.96	−1.83	112.06	−2.09	102.85	−2
Chloramphenicol	150.5	−2.11	1.15	−1.91	183.22	−2.23	178.91	−2.19	162.48	−2.02
Flufenamic acid	134	−5.35	4.6	−5.19	178.95	−5.64	165.02	−5.5	183.83	−5.69
Griseofulvin	219	−3.25	2.71	−4.15	139.13	−3.35	127.36	−3.23	115.41	−3.11
Hydrochlorothiazide	269	−2.69	−0.16	−1.78	317.53	−2.27	190.79	−1	186.21	−0.95
Nalidixic acid	229	−3.61	1.27	−2.81	203.54	−2.56	167.32	−2.19	198.89	−2.51
Nicotinic acid	237.5	−0.85	0.29	−1.92	110.59	−0.65	135.54	−0.9	167.72	−1.22
Papaverine	146.5	−3.87	4.19	−4.91	113.66	−4.58	138.88	−4.83	124.55	−4.69
Perylene	278	−8.8	6.34	−8.37	151.23	−7.1	166	−7.25	169.45	−7.28
Pyrene	150	−6.18	5.19	−5.94	108.61	−5.53	131.49	−5.75	102.86	−5.47
Quinidine	170	−2.81	2.82	−3.77	122.24	−3.29	158.15	−3.65	121.73	−3.29
Salicylamide	140	−1.84	0.74	−1.39	148.96	−1.48	161.31	−1.6	177.31	−1.76
Salicylic acid	159	−1.94	1.96	−2.8	126.83	−2.48	151.54	−2.73	151.97	−2.73
Sulfacetamide	183	−1.51	0.15	−1.23	188.22	−1.28	155.26	−0.95	223.86	−1.64
Sulfadiazine	254.5	−2.73	0.25	−2.05	209.12	−1.59	168.86	−1.19	206.35	−1.56
Sulfamethazine	199.5	−2.73	0.43	−1.68	206.56	−1.75	187.3	−1.55	210.09	−1.78
Sulfanilamide	165.5	−1.36	−0.16	−0.75	161.27	−0.7	153.04	−0.62	127.98	−0.37
Thymine	316.5	−1.5	−0.8	−1.62	154.44	0.01	188.64	−0.34	254.67	−1
Thymol	50.5	−2.19	3.16	−2.92	22.66	−2.64	44.09	−2.85	24.33	−2.65
Tolbutamide	129	−3.47	2.04	−2.58	140.45	−2.69	144.96	−2.74	161.97	−2.91
Triphenylene	196.5	−6.73	5.77	−6.99	127.13	−6.29	159.55	−6.62	182.53	−6.85
Uracil	330	−1.49	−1.28	−1.27	158.73	0.44	185.28	0.18	229.04	−0.26
RMSE				0.77		0.86		0.95		0.93
R^2				0.84		0.83		0.8		0.8
σ		1.95		0.84		0.8		0.88		0.89
Bias				−0.06		0.31		0.36		0.27

to see a more significant improvement in the present article, going from predicted to experimental input, as we must consider the total prediction errors over MP and logP and that inherent to the GSE. Palmer *et al.* are testing the error attributable to the MLR equation alone. Additionally, the GSE parameters are arrived at by thermodynamic consideration, not MLR fitting, i.e. the GSE is not trained.[2b,4a]

Using the revised GSE, the bias reduces to almost zero. This provides evidence for there being little systematic error in the model. In turn this also provides reassuring evidence that the AlogP predictions are of good accuracy. Given that the AAE of the SOL-30 MP prediction is ~50 °C, which equates to 0.5 logS units, one would expect to see the AAE in the solubility predictions drop by ~0.5 logS units when experimental melting points are used. We indeed see this, further suggesting the model has little systematic error, see Table 3 and supporting information, Table S2. This adds further support for the idea that, in this case, the model is systematically improved by providing improved data.

Tables 2 and 3 and Figures 7–10 present a body of evidence showing that logS predictions of the SOL-30 dataset are good, generally close to the 1 logS accuracy target. Despite SOL-30 being a small dataset, the results are in agreement with work carried out by others. Previous results show the predictive ability of the GSE to be good.[2b,4a,26] This work confirms the predictive ability of the GSE and goes further showing that one can achieve good predictions of solubility despite having fairly low quality melting point data. The above data show that fairly imprecise predictions of melting points are enough (at least on this small dataset), to produce accurate and useful logS predictions. All prediction RMSEs fall within the standard deviation of the experimental data (1.95 logS units) and therefore, we considered these predictions to be useful. Further to this, when the revised GSE is employed the models meet the 1 logS unit RMSE target. In this case the best model is that of the revised GSE using PLS predicted melting points (RMSE = 0.86 R2 = 0.83).

An interesting note is that on occasion the absolute errors in the predicted melting points can be in excess of

Table 3. The absolute differences between the experimental and predicted melting points from PLS, RF and SVM.

Molecule name	PLS (|exp-pred|)	RF (|exp-pred|)	SVM (|exp-pred|)
1,3,5−trichlorobenzene	35.2	26.54	21.11
1−Naphthol	20.7	57.34	34.2
4−Aminobenzoic acid	42.9	39.84	19.51
5,5−Diphenylhydantoin	71.55	121.3	104.37
Acetanilide	39.19	54.58	51.05
Adenosine	5.39	53.95	119.59
Antipyrine	10.1	33.18	17.8
Benzamide	33.66	37.76	44.5
Benzoic acid	36.54	10.44	19.65
Chloramphenicol	32.72	28.41	11.98
Flufenamic acid	44.95	31.02	49.83
Griseofulvin	79.87	91.64	103.59
Hydrochlorothiazide	48.53	78.21	82.79
Nalidixic acid	25.46	61.68	30.11
Nicotinic acid	126.91	101.96	69.78
Papaverine	32.84	7.62	21.95
Perylene	126.77	112	108.55
Pyrene	41.39	18.51	47.14
Quinidine	47.76	11.85	48.27
Salicylamide	8.96	21.31	37.31
Salicylic acid	32.17	7.46	7.03
Sulfacetamide	5.22	27.74	40.86
Sulfadiazine	45.38	85.64	48.15
Sulfamethazine	7.06	12.2	10.59
Sulfanilamide	4.23	12.46	37.52
Thymine	162.06	127.86	61.83
Thymol	27.84	6.41	26.17
Tolbutamide	11.45	15.96	32.97
Triphenylene	69.37	36.95	13.97
Uracil	171.27	144.72	100.96
Average	48.25	49.22	47.44

100 °C (see Table 3), yet accurate predictions of solubility are still achievable. Thus, the GSE can be considered robust, even if provided with very poor melting point data. This is likely due to the logP term being the dominant term in the GSE and hence the solubility predictions. The melting point prediction errors of 'brick dust' (low solubility) compounds will have a larger impact as the melting points of these compounds tend to be more difficult to predict (Figure 6). The GSE scales the empirical input so that only 1% of the predicted melting point value minus 25 °C (0.01x (MP-25)) enters the solubility prediction (see Equation 1), with each 1 °C change in melting point giving a change of 0.01 in the predicted logS.

We can now offer an answer to our second question ('can values of this level of accuracy be usefully applied to predicting aqueous solubility?'). This answer is simply "yes". We have seen that even with very bad predictions of the melting point (RMSE > 100 °C), the accuracy of the solubility prediction is still very high, achieving the 1 logS criterion for the SOL-30 dataset. This is for a small dataset and thus this finding needs to be confirmed on a larger dataset.

3.4 Models of logP and their effect on the GSE

Although we have answered the two questions we initially set out to answer, there remains an untested aspect, the logP prediction. Here we test the dependency of the GSE on the chosen model of logP prediction. Figures 11–14

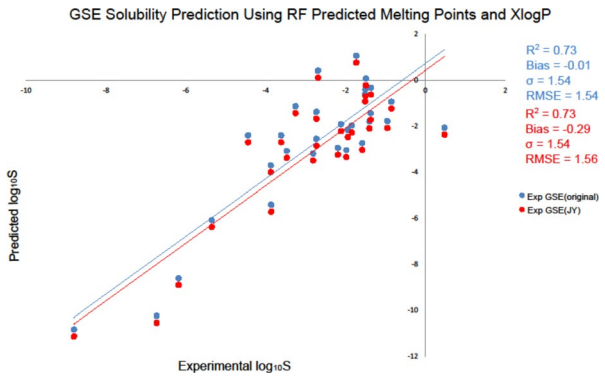

Figure 11. A prediction of solubility for the SOL-30 molecules using the General Solubility Equation with melting points from RF and predicted logP from XlogP. Bias, standard deviation (SD) and RMSE are quoted in logS units.

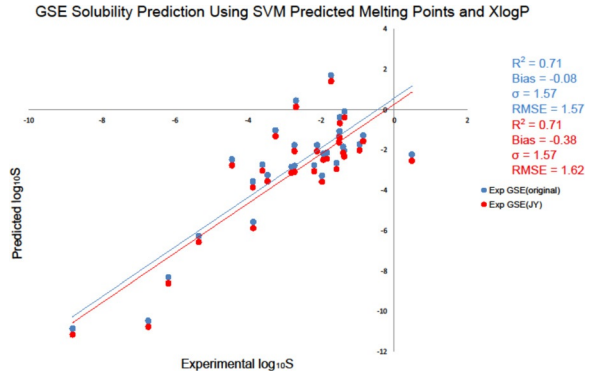

Figure 12. A prediction of solubility for the SOL-30 molecules using the General Solubility Equation with melting points from SVM and predicted logP from XlogP. Bias, standard deviation (SD) and RMSE are quoted in logS units.

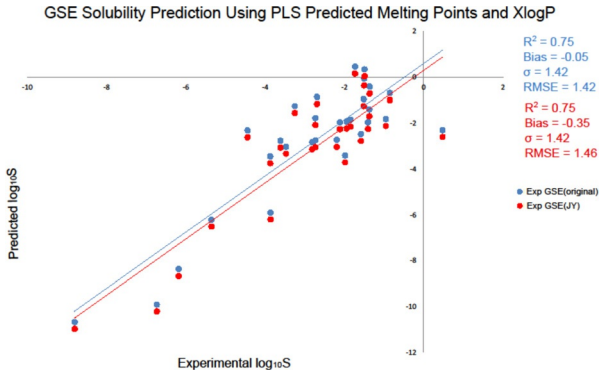

Figure 13. A prediction of solubility for the SOL-30 molecules using the General Solubility Equation with melting points from PLS and predicted logP from XlogP. Bias, standard deviation (SD) and RMSE are quoted in logS units.

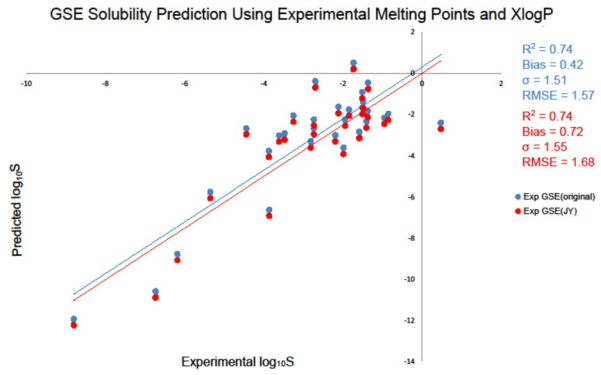

Figure 14. A prediction of solubility for the SOL-30 molecules using the General Solubility Equation with melting points from experiment and predicted logP from XlogP. Bias, standard deviation (SD) and RMSE are quoted in logS units.

show the solubilities predicted using logP values obtained from the XlogP algorithm instead of the AlogP algorithm.

Figures 11–14 demonstrate a reversal in the optimum GSE version; they now favour the original GSE. There is a large increase in the systematic error from the original to the revised GSE leading to a deteriorating RMSE value. The RMSE values found when using the XlogP algorithm to predict the logP values, are considerably worse than those found when the AlogP algorithm is used. All other aspects of the models are equivalent. In addition, a non-intuitive result is found on substituting predicted with experimental melting points: the RMSE increases suggesting a worse prediction.

There appears to be some volatility in predictive accuracy from the GSE dependent on the source of the empirical parameters. Based on the data in this study, the GSE accuracy seems more dependent on the source of the logP data than on the melting point data. We note that any change in the logP value results in an essentially equivalent change in the logS predicted by the GSE. This study shows models which make accurate predictions of solubility from imprecise predictions of melting point and the well-known AlogP method of logP prediction. Additionally, due to the logP term's dominance in the GSE, the GSE is more sensitive to the choice of algorithm for logP prediction than melting point prediction.

4 Conclusion

From this work we find several conclusions. In answer to our first question, we find that it is beyond the capabilities of simple machine learning models, utilising a modest number of 2D chemical descriptors, to predict chemically accurate melting points. However, it is possible to make useful predictions and predictions of qualitative use. These predictions can be considered useful, depending on the end goal of making such predictions, i.e. they can be considered useful here if the end goal is solubility prediction, as they enable the use of an accurate solubility prediction scheme, the GSE.

We can conclude in answer to the second question that, "yes", values of melting point prediction with RMSE of tens of degrees Celsius can make a useful and accurate prediction of aqueous solubility via the GSE. We note that a variation of 50 °C in melting point leads to only a 0.5 logS unit change in the predicted solubility. We present results of the application of melting point predictions of this level of accuracy to solubility predictions via the GSE. We additionally conclude that the GSE is more reliant on the accuracy of logP predictions than on the level of accuracy of melting point data it is provided with.

Conflict of Interest

None declared.

Acknowledgements

JLMcD and JBOM would like to thank SULSA for funding. The authors gratefully acknowledge the use of the EaStChem research computing facility and Dr. Herbert Früchtl for its maintenance. JLMcD would like to thank Prof. Graeme Day, Prof. Michael Bühl, Dr. David Palmer and Ms Rachael Skyner for useful discussions.

References

[1] A. Carletta, C. Meinguet, J. Wouters, A. Tilborg, *Cryst. Growth. Des.* **2015**, *15*.
[2] a) N. Jain, S. H. Yalkowsky, *J. Pharm. Sci.* **2001**, *90*, 234–252; b) Y. Ran, N. Jain, S. H. Yalkowsky, *J. Chem. Inf. Comp. Sci.* **2001**, *41*, 1208–1217; c) S. H. Yalkowsky, S. C. Valvani, *J. Pharm. Sci.* **1980**, *69*, 912–922.
[3] R. E. Skyner, J. L. McDonagh, C. R. Groom, T. van Mourik, J. B. O. Mitchell, *PCCP* **2015**, *17*, 6174–6191.
[4] a) Y. Ran, S. H. Yalkowsky, *J. Chem. Inf. Comp. Sci.* **2001**, *41*, 354–357; b) T. Sanghvi, N. Jain, G. Yang, S. H. Yalkowsky, *QSAR Comb. Sci.* **2003**, *22*, 258–262.
[5] a) A. K. Ghose, G. M. Crippen, *J. Comput. Chem.* **1986**, *7*, 565–577; b) A. K. Ghose, G. M. Crippen, *J. Chem. Inf. Comp. Sci.* **1987**, *27*, 21–35; c) V. N. Viswanadhan, A. K. Ghose, G. R. Revankar, R. K. Robins, *J. Chem. Inf. Comp. Sci.* **1989**, *29*, 163–172.
[6] R. Wang, Y. Fu, L. Lai, *J. Chem. Inf. Comp. Sci.* **1997**, *37*, 615–621.
[7] a) C. Hansch, A. Leo, D. Hoekman, S. R. Heller, *Exploring Qsar*, American Chemical Society Washington, DC, **1995**; b) A. Leo, P. Y. C. Jow, C. Silipo, C. Hansch, *J. Med. Chem.* **1975**, *18*, 865–868; c) A. J. Leo, *Chem. Rev.* **1993**, *93*, 1281–1306.
[8] a) L. D. Hughes, D. S. Palmer, F. Nigsch, J. B. O. Mitchell, *J. Chem. Inf. Model.* **2008**, *48*, 220–232; b) C. A. S. Bergstrom, U. Norinder, K. Luthman, P. Artursson, *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 1177–1185; c) F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch, J. B. O. Mitchell, *J. Chem. Inf. Model.* **2006**, *46*, 2412–2422.
[9] I. V. Tetko, Y. Sushko, S. Novotarskyi, L. Patiny, I. Kondratov, A. E. Petrenko, L. Charochkina, A. M. Asiri, *J. Chem. Inf. Model.*, *54*, 3320–3329.
[10] W. Kew, J. B. O. Mitchell, *Mol. Inf.* **2015**, In Press. DOI: 10.1002/minf.201400122
[11] U. P. Preiss, W. Beichel, A. M. T. Erle, Y. U. Paulechka, I. Krossing, *ChemPhysChem* **2011**, *12*, 2959–2972.
[12] J. C. Bradley, A. Lang, A. Williams, online - http://datahub.io/dataset/open-melting-point-data.
[13] J. L. McDonagh, N. Nath, L. De Ferrari, T. van Mourik, J. B. O. Mitchell, *J. Chem. Inf. Model.* **2014**, *54*, 844–856.
[14] A. K. Ghose, V. N. Viswanadhan, J. J. Wendoloski, *J. Phys. Chem. A* **1998**, *102*, 3762–3772.
[15] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 1947–1958.
[16] D. Basak, S. Pal, D. C. Patranabis, *Neu. Inf. Pro.- Letters and Reviews* **2007**, *11*, 203–224.
[17] H. Abdi, *Encyclopedia for research methods for the social sciences* **2003**, 792–795.
[18] a) J. B. O. Mitchell, *Future Med Chem* **2011**, *3*, 451–467; b) J. B. O. Mitchell, *WIRES Comput. Mol. Sci.* **2014**, 468–481; c) A. R. Leach, V. J. Gillet, *An introduction to chemoinformatics*, Springer, **2007**.
[19] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, E. Willighagen, *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 493–500.
[20] 27/05/2015, OCHEM – Wiki, CDK molecular descriptors, url http://wiki.qspr-thesaurus.eu/w/CDK.
[21] a) Y. A. Abramov, *Mol. Pharm.* **2015**, *12*, 2126–2141; b) D. S. Palmer, N. M. O'Boyle, R. C. Glen, J. B. O. Mitchell, *J. Chem. Inf. Model.* **2006**, *47*, 150–158.
[22] M. Karthikeyan, R. C. Glen, A. Bender, *J. Chem. Inf. Model.* **2005**, *45*, 581–590.
[23] D. S. Palmer, J. L. McDonagh, J. B. O. Mitchell, T. van Mourik, M. V. Fedorov, *J. Chem. Theory Comput.* **2012**, 3322–3337.
[24] M. Stuart, K. Box, *Anal. Chem.* **2005**, *77*, 983–990.
[25] D. S. Palmer, J. B. O. Mitchell, *Mol. Pharm.* **2014**, *11*, 2962–2972.
[26] G. Yang, Y. Ran, S. H. Yalkowsky, *J. Pharm. Sci.* **2002**, *91*, 517–533.

Received: May 6, 2015
Accepted: June 5, 2015
Published online: July 20, 2015

