# Are the Sublimation Thermodynamics of Organic Molecules Predictable?

*James L. McDonagh [a,c], David S. Palmer [b], Tanja van Mourik [c], and John B.O. Mitchell [c*]*

a Manchester Institute of Biotechnology, The University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK

b Department of Pure and Applied Chemistry, University of Strathclyde, Thomas Graham Building, 295 Cathedral Street, Glasgow, Scotland, United Kingdom, G1 1XL

c School of Chemistry, University of St Andrews, North Haugh, St Andrews, Fife, Scotland, United Kingdom, KY16 9ST

**E-mail jbom@st-andrews.ac.uk*

ABSTRACT

We compare a range of computational methods for the prediction of sublimation thermodynamics (enthalpy, entropy and free energy of sublimation). These include a model from theoretical chemistry that utilizes crystal lattice energy minimization (with the DMACRYS program) and QSPR models generated by both machine learning (Random Forest and Support Vector Machines) and regression (Partial Least Squares) methods. Using these methods we investigate the predictability of the enthalpy, entropy and free energy of sublimation, with consideration of whether such a method may be able to improve solubility prediction schemes. Previous work has suggested that the major source of error in solubility prediction schemes involving a thermodynamic cycle via the solid state is in the modeling of the free energy change away from the solid state. Yet contrary to this conclusion other work has found that the inclusion of terms such as the enthalpy of sublimation in QSPR methods

does not improve the predictions of solubility. We suggest the use of theoretical chemistry terms, detailed explicitly in the methods section, as descriptors for the prediction of the enthalpy and free energy of sublimation. A dataset of 158 molecules with experimental sublimation thermodynamics values and some CSD refcodes has been collected from the literature and is provided with their original source references.

## Introduction

Sublimation thermodynamics has not enjoyed the attention received by its solvation/hydration and solution counterparts. This potentially stems from the experimental difficulties in determining thermodynamic values for sublimation.[1] Historically, often only the enthalpy of sublimation was determined as the free energy was not accessible from the calorimetric measurements used.[2] It has been estimated that there are 1.8 times as many enthalpies as free energies identified from experiment in the literature.[2] The data available in the literature have been determined by a variety of challenging experimental techniques,[2,3] making it difficult to collect a reliable set in which the experimental noise is minimized.[3] The entropy of sublimation is often back calculated from the enthalpy, free energy and temperature;[2,4,5] this potentially makes error margins difficult to assess. *In silico* Quantitative Structure Property Relationship (QSPR) studies of solubility often lack an explicit account of sublimation thermodynamics in relation to solubility.[6] Sublimation thermodynamics are essential in determining the strength of solid-state intermolecular interactions. Hence, sublimation thermodynamics have important effects in diverse industries including: dyes, agrochemicals, environmental contaminants, and pharmaceuticals (in determining solubility of drug candidates), to name but a few.[1,7–10] As solubility modeling becomes increasingly widespread, with greater accuracy demands, there is a requirement for wider study of sublimation thermodynamics, as the free energy of sublimation is one of four the components in the two most commonly used thermodynamic cycles for solubility prediction. These

thermodynamic cycles are depicted in **Figure 1.** In the sublimation cycle (top of **Figure 1**), the free energy of solution ($\Delta G_{sol}$) is computed as the sum of the free energy of sublimation ($\Delta G_{sub}$) (solid to gas transition) and free energy of hydration/solvation ($\Delta G_{hyd}/\Delta G_{sol}$) (gas to solution). In the fusion cycle (bottom of **Figure 1**) the free energy of solution is computed as the sum of the free energy of fusion ($\Delta G_{fus}$) (solid to supercooled transition) and free energy of mixing ($\Delta G_{mix}$) (supercooled to solution transition). In previous work, we have demonstrated that the free energy of hydration is predictable to a good level of accuracy using theoretical chemistry calculations.[11] In the same work, we found that the largest single error in the prediction of the solution free energy could be attributed to the sublimation free energy calculation. In recent work, Docherty *et al*.[1] have demonstrated a deconvolution of the effects of crystal packing and solvation on solubility, enabling a thorough analysis of the contributions to poor solubility of drug-like molecules. Salahinejad *et al*.[8] have recently provided a novel QSPR model for predicting the enthalpy of sublimation with good accuracy using Volsurf and CPSA descriptors. Abramov[9] has shown that the major source of error in solubility prediction via the fusion cycle can be attributed to the fusion step in QSPR models. It is suggested by Abramov that this may be due to a lack of suitable descriptors for the solid state. Salahinejad *et al*. have additionally provided a QSPR model for the prediction of aqueous solubility, which utilises two descriptors from physical chemistry, lattice energy and enthalpy of sublimation. Salahinejad *et al*. concluded that these descriptors did not improve the predictions of the models.[7]

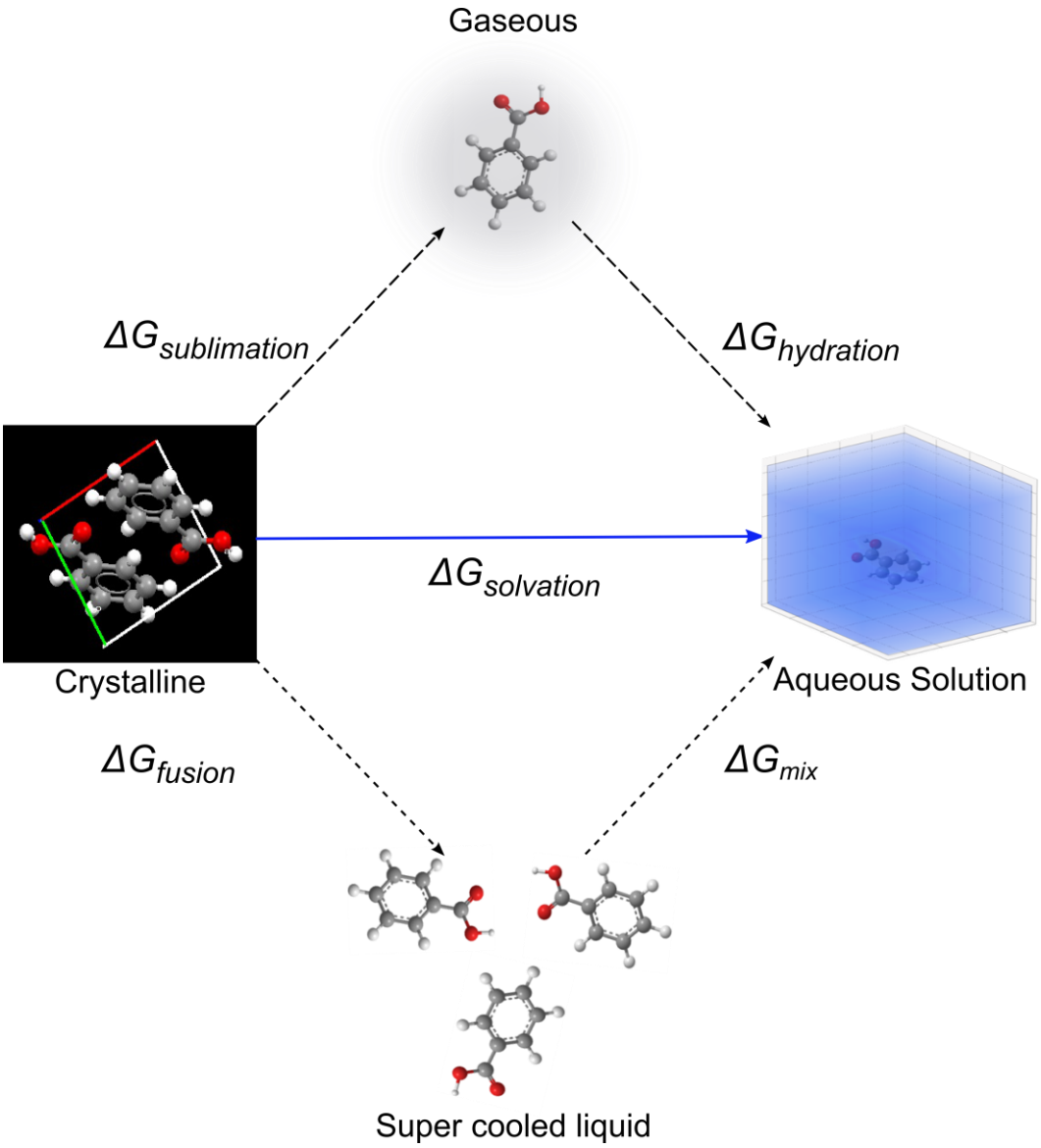



**Figure 1**. Two thermodynamic cycles commonly used in solubility prediction. The first, the sublimation cycle, calculates the free energy of solution ($\Delta G_{sol}$) by summing the calculated free energy of sublimation ($\Delta G_{sub}$) and free energy of hydration ($\Delta G_{hyd}$). The second, the fusion cycle, calculates the free energy of solution by summing the calculated free energy of fusion ($\Delta G_{fus}$) and free energy of mixing ($\Delta G_{mix}$).

One major research area where the thermodynamic cycles shown in **Figure 1** are employed is in the complex multi-step drug discovery process, where experimental and predictive scientific methodologies need to be applied. The cycles themselves can be disguised by the calculation method; for example, the equations of the General Solubility Equation are based on the fusion cycle.[12] Thermodynamic intrinsic aqueous solubility (the solubility of an unionised species in a saturated solution)[13,14] is a key parameter in the process, due to its

determining contribution to bio-availability and its use in determining pH-dependent solubility. Thermodynamic solubility is a parameter considered throughout the drug discovery process, from lead optimization to formulation. In recent years, it has been suggested that up to 40% of new drug molecules are effectively insoluble[15] and up to 75% are classified as having low solubility by the Biopharmaceutics Classification System (BCS).[16,9]

At the lead optimization stage, it is not practical to carry out experimental determinations of thermodynamic solubility on large libraries of compounds. As a result, computational methods are employed to make predictions of numerous properties including solubility.[9,17,18] There are now numerous methods to predict solubility, with most varying in how the solvation/hydration step is modeled. Many of these solvation/hydration methodologies have been summarized in a recent review by Skyner *et al.*[19] Commonly, QSPR methods are employed due to their speed, convenience and accuracy, when provided with a suitable training dataset.[10,20–22,23] These methods represent the current state-of-the-art in practical solubility prediction. However, QSPR methods lack the theoretical basis of a fundamental physical theory, hence limiting their interpretability and the understanding that can be gained from their use. Recently, the standard dogma attached to these models, which stated that their accuracy was limited by the accuracy of the experimental data, has been challenged by Palmer and Mitchell,[24] suggesting there is a limit inherent to current QSPR methodologies. Even more recently, it was demonstrated that such models have difficulty representing the state change from crystalline to liquid (fusion/melting process), possibly due to a lack of suitable descriptors for the solid state.[9] The creation of standard experimental datasets for solubility prediction is on-going.[25]

There are methods which apply physically motivated fitted equations[12] or full calculations from first principles[11,26] to solubility prediction. The fitted equation approach has shown good accuracy, but is limited by requirements for additional experimental input, although

promising attempts have been made to predict some of these quantities.[20,27–30] There are examples in the literature which utilize predicted melting points and logP (octanol water partition coefficient) values for solubility predictions via the general solubility equation.[20,30,31] The first principles calculation methods have generally been less accurate and more time consuming, but can provide more fundamental understanding of the process via physically meaningful decomposition of the predicted solution free energy. Recently, much work has focused on improving first principles calculation methods of hydration/solvation free energy and has resulted in much improved methods.[19,32–34] Generally, full calculation methods for solubility prediction either use directly, or make reference to, one of the two thermodynamic cycles in **Figure 1**. We recently presented work showing that these methods, whilst not matching the accuracy of QSPR models, are now capable of reasonable predictions in an acceptable amount of CPU time.[11]

In this work, we focus on addressing the issues of sublimation free energy prediction and its incorporation into such thermodynamic cycles. Our recent work showed that the majority of the solubility prediction error via a sublimation cycle originated from the sublimation calculation,[11] which is in agreement with recent publications.[9] We present predictions of the enthalpy ($\Delta H_{sub}$), entropy ($\Delta S_{sub}$) and free energy of sublimation ($\Delta G_{sub}$) using the following:

- Theoretical chemistry calculations
- QSPR models based on conventional 2D descriptors only
- QSPR models based on predicted values from theoretical chemistry only
- QSPR models based on both 2D descriptors and predicted values from theoretical chemistry

Our hypothesis is that the predictions from theoretical chemistry will be better, physically meaningful, descriptors for the solid state. The relevant terms are predicted from theoretical chemistry using the programs DMACRYS[35], GDMA2[36,37] and Gaussian 09 (G09).[38]

## Methods

*Dataset*

Sublimation data for a diverse range of organic crystals were found by searching the appropriate literature for data at 298 - 298.15 K that met our criteria:

1. For each molecule experimental values for the enthalpy, entropy and free energy of sublimation must be available or calculable from a single literature source.
2. A crystal structure should be available in the Cambridge Structural Database (CSD).[39]
3. The literature should define the polymorphic state or pseudo-polymorphic state.

Using the first criterion, a total of 158 molecules were identified (SUB-158 dataset). The second criterion caused the removal of 62 molecules (either the structure was not present or unusable in the current workflow), hence leaving a 96-molecule dataset. If we apply the final criterion the dataset would reduce to four molecules. Clearly this would be insufficient for our needs, but does highlight a lack of sublimation data with polymorphic information in the literature. When descriptors dependent upon polymorphic information are used, such as those from theoretical chemistry, it is very important that the experimental data are for the same polymorphic form. This is desired as different polymorphs can present very different physicochemical properties such as enthalpy of sublimation and solubility. Hence, if inconsistent experimental data are used in QSPR model training and descriptor generation, then very poor agreement may be obtained.[40] In the present case, all of the theoretical chemistry terms are dependent on polymorphic information. Thus, in order to keep our

dataset at a sufficient size, we did not apply the final criterion in its original form. Instead, we opted to minimize, using DMACRYS,[35] the lattice energy of all of the polymorphs reported in the CSD for each molecule in the dataset. We worked with the assumption that the polymorph with the lowest lattice energy was the most stable and therefore the form whose sublimation thermodynamics were reported. Whilst some may consider this assumption crude, we have applied the assumption previously and found promising results; it would be of interest to attempt similar analysis with improved data to determine whether such an assumption is generally reasonable. Due to the time-consuming nature of this approach, we chose a dataset of 60 molecules selected on the basis of the best available crystal structures. Minimization problems with some of the polymorphs meant that the 60 molecule dataset was eventually reduced to 48 compounds, which will be referred to as the SUB-48 dataset. In this work we utilize the SUB-48 dataset to test a variety of prediction methods. We utilise the SUB-158 dataset as a larger QSPR prediction dataset assessing the use of standard, computationally efficient, 2D descriptors for such predictions. Supporting Information **Table S1** and **Table S2** give the datasets, CSD refcodes and the SMILES structures used for the descriptor generation.

*Theoretical Chemistry Methods*

DMACRYS[35] is a periodic lattice simulation program, capable of efficiently minimizing crystal structures within the rigid body approximation to a local minimum. DMACRYS uses distributed multipoles to accurately represent the electrostatic interactions. The multipoles are calculated by distributed multipole analysis[36] in GDMA2 based on the density matrix from a prior quantum chemical calculation using Gaussian 09[38] at the B3LYP/6-31G** level of theory.[41] An empirical Buckingham potential is applied to account for repulsion and dispersion. The potential parameters are those from the FIT potential,[42,43,44] and are provided in the Supporting Information of reference 1. DMACRYS directly yields calculated lattice

energies and is also capable of approximately assessing the crystalline entropy from the phonon modes. DMACRYS calculates the contribution from the optical phonons at the gamma point (k (Bloch wave vector) = 0) where the acoustic modes decay to 0. A hybrid Debye-Einstein approximation is applied to account for acoustic and optical phonon modes away from the gamma point.[35] This approximation enables the calculation of the phonon density of states. The entropy is finally calculated as the negative of the partial derivative of the Helmholtz free energy with respect to temperature at constant volume.[45] Vacuum (approximating the gas phase) calculations are carried out using Gaussian 09. A single molecule is extracted from the crystal and geometry optimized at the B3LYP/6-31G** level of theory to the vacuum local minimum. The entropy of the isolated molecule is evaluated considering the statistical thermodynamics of an ideal gas using the routines available in Gaussian 09.[46]

From these data we can calculate approximations for the enthalpy, entropy and free energy of sublimation using the following equations:

$$\Delta H_{sub} = -U_{latt} - 2RT$$

**Equation 1**.

$\Delta H_{sub}$ is the enthalpy of sublimation, $U_{latt}$ is the lattice energy, R is the gas constant and T is the temperature in Kelvin. **Equation 1** was originally given by Gavezzotti and Filippini[47] and is arrived at on the following theoretical assumptions: The enthalpy is defined as H = U + pV. Since for one mole of an ideal gas pV = RT, we approximate the environmental contributions by RT. The equipartition theorem is then applied to calculate the contributions from molecular motion. Gaseous rotations and translations provide 1½RT each per mole, hence 3RT in total. Crystal lattice phonon (collective lattice vibrations and rotations) contributions contain both kinetic and potential energy terms for six degrees of freedom per

molecule (three coordinates and three Euler angles) and thus provide 2 x 6 x ½RT = 6RT. Intramolecular vibrations are assumed to be the same in each phase and hence assumed to cancel out. Summing these contributions, and noting that for one mole of ideal gas we can simply write $\Delta H = \Delta U + pV = \Delta U + RT$, which leads to $\Delta H_{sub} = (-U_{latt} + RT) - 6RT + 3RT = -U_{latt} - 2RT$. In this way, the -2RT contribution to the enthalpy is derived from consideration of molecular and crystal degrees of freedom. $U_{latt}$ describes the energy of assembling the lattice from infinitely separated molecules, in essence the reverse of sublimation, and hence the equation for $\Delta H_{sub}$ contains a $-U_{latt}$ term. The sublimation entropy is given by:

$$\Delta S_{sub} = S_{crys} - (S_{gas}^{trans} + S_{gas}^{rot})$$

**Equation 2**.

$\Delta S_{sub}$ is the enthalpy of sublimation, $S_{crys}$ is the crystalline phonon entropy, $S_{gas}^{trans}$ is the gaseous translational entropy contribution and $S_{gas}^{rot}$ is the gaseous rotational entropy contribution. We assume in **Equation 2** that there is no change in the electronic entropy across the phase transition. We also assume the intermolecular and intramolecular vibrational and rotational contributions are decoupled in the crystal; there is therefore no net change in the intramolecular vibrational entropy over the phase transition.

Finally, the free energy is given by the usual Gibbs equation (**Equation 3**)

$$\Delta G_{sub} = \Delta H_{sub} - T\Delta S_{sub}$$

**Equation 3**.

*Cheminformatics Descriptors SUB-48*

Descriptors were calculated from a SMILES[48] representation of the molecules using the open source Java library, the Chemistry Development Kit (CDK).[49–51] 132 descriptors were found to contribute information, i.e. their variance was not zero, for the SUB-48 dataset. Additionally, to avoid the use of highly correlated descriptors, an inter-descriptor Pearson correlation threshold 0.9 was applied to remove highly correlated descriptors. Auto-scaling (subtracting the mean and dividing by the standard deviation) was applied to all remaining descriptors. Three descriptor sets were used for each thermodynamic term:

1. Theoretical chemistry values
2. CDK descriptors
3. Theoretical chemistry values and CDK descriptors.

A list of all descriptors is present in the Supporting Information (**Table S3)**. For clarity, the theoretical chemistry descriptors used are as follows: the predicted ΔH of sublimation, predicted ΔS of sublimation, predicted ΔG of sublimation, phonon entropy, gas phase rotational entropy and finally gas phase translational entropy. This is a total of six descriptors.

*QSPR Descriptors SUB-158*

X-ray structures were not always available or suitable for all molecules in SUB-158[52–105]; the theoretical chemistry terms are therefore inaccessible for some of the molecules. The predictions of the sublimation thermodynamics of SUB-158 are therefore carried out with only CDK descriptors. A Pearson correlation threshold of 0.9 and auto-scaling were applied over all descriptor sets for SUB-158 dataset.

*QSPR models*

In this work, the QSPR models were generated from the descriptor sets by three methods: Random Forest (RF), Support Vector Machines (SVM) and Partial Least Squares (PLS).

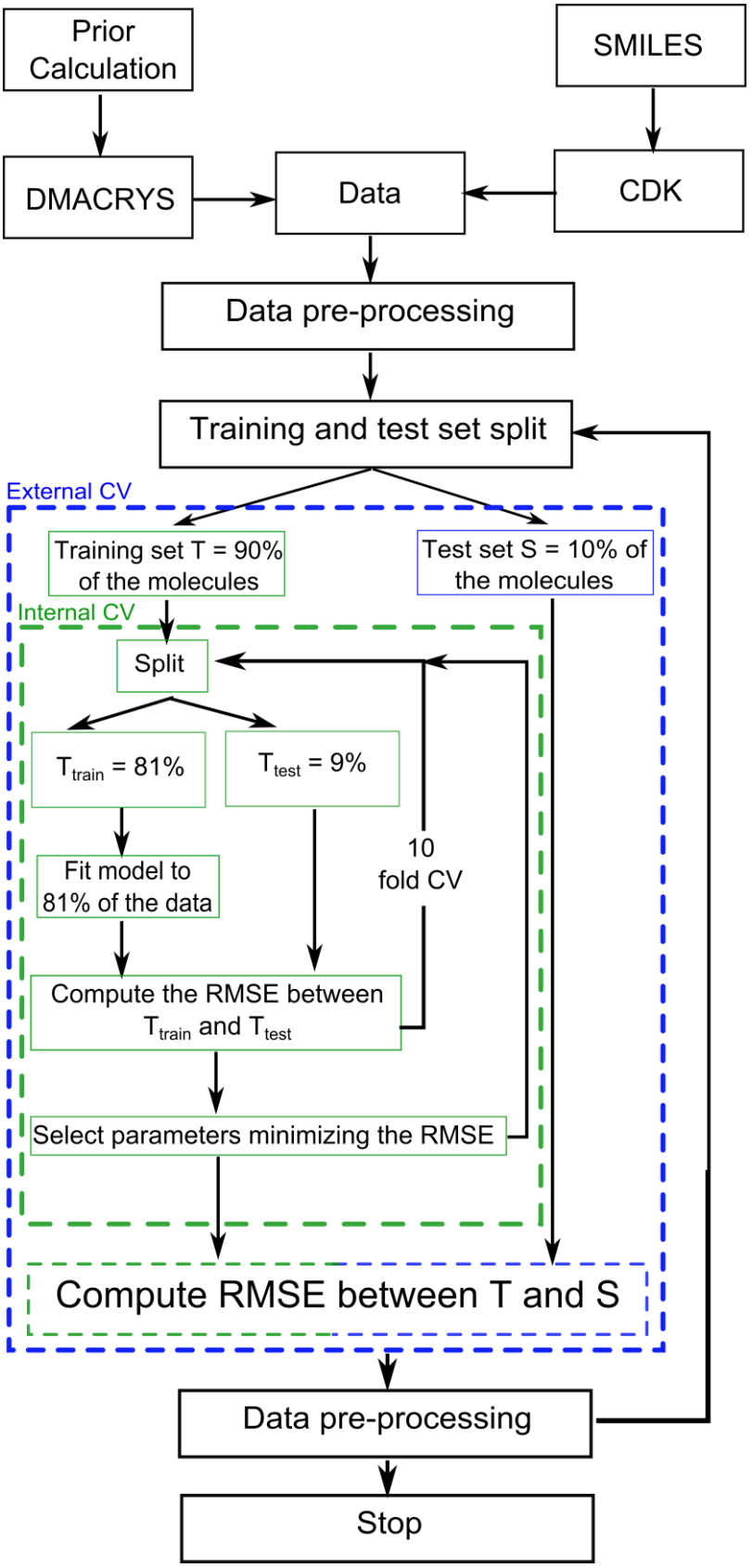
These methods have been previously discussed in a number of publications[10,19,27,106] and are only briefly outlined below. The workflow is outlined in **Scheme 1**.

RF is an ensemble learning method that generates a forest of decision trees applicable to regression and classification tasks. Here we apply RF to a regression task. The method follows a general workflow of selecting, with replacement, a random sample from the training molecules. This sample is then in turn used to grow a regression tree to its maximum extent (as given by a parameter called *nodesize* that gives the number of data points in each node below which that node is not further subdivided), by calculating the best split available to the algorithm from a random subset of the descriptors. These steps are repeated until a defined number of trees are created.[107] This eventually leads to a forest of regression trees. The RF prediction is the average prediction over all of the trees.

SVM seeks an optimal regression function by projecting the features into a higher order feature-space. A parameter ε is selected, which quantifies a margin of acceptable error from the regression in the feature space. The SVM function then attempts to predict the *y* responses to the *x* input variables within the defined margin. A penalty is applied for any prediction that lies outside the margin. Additionally, the regression function is required to remain as flat as possible, to avoid over-fitting the function.[108]

PLS is a regression method, which reduces the number of descriptors used by combining descriptors to form linear combinations based on their relative explanatory ability. These linear combinations are called latent variables. The method also attempts to maximally explain the co-variance between the latent variables and the *y* independent variable.[108] As the number of latent variables is far less than the number of descriptors, the method protects against over fitting.

The QSPR models were generated using stacked ten-fold cross validations for training and testing, i.e. an external ten-fold cross validation making a random split of all of the data for training (90%) and testing (10%). Thus, we create ten separate models for each modeling method built from 90% of the data and tested on the remaining 10%, such that each instance is in the test set for one model, hence providing a prediction for each data point. The test set results for each fold is provided in the Supporting Information. An internal ten-fold cross validation (81%:9% of the original data) then optimizes the model parameters during training. In PLS we optimize the number of components (*ncomp*), which ranges between 1 and 20. In RF, the number of descriptors in the random subset (*mtry*) is optimized (values range between 2 and 137 via grid searching in R's train function), whilst the number of trees was set to a constant value of 1000. Finally, in the SVM model where a radial basis kernel is used, we optimize the cost parameter (*C*) (range of values 0.25 – 131072.00, with each step doubling the previous value) whilst the kernel (radial basis) parameter σ is selected using R's train function and the epsilon loss parameter is set to 0.1.[109,110] We use the RMSE as a fitness parameter. This is an efficient method to produce multiple QSPR models from different machine learning algorithms. The code is available from the corresponding author's website[111], GitHub[112] and the Supporting Information. We have previously applied similar methodologies to full solubility prediction with some success.[10] The overall workflow procedure is as follows:




**Scheme 1**. QSPR model generation scheme including descriptor set generation.

*Statistics*

The following statistics were used to analyze the predictions. Respectively, **Equations 4 – 7** are: correlation coefficient ($R^2$), the Root Mean Square Error (RMSE), the standard deviation of the prediction error (σ) and the bias.

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

**Equation 4.**

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

**Equation 5.**

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

**Equation 6.**

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

**Equation 7.**

where $y_{exp}^{i}$ is the experimental value for molecule *i*, $y_{pred}^{i}$ is its predicted value, $y_{exp-pred}^{i}$ is the difference between the experimental and predicted values, N is the number of data points and $\bar{y}$ is the mean. A sample standard deviation is used to calculate the experimental standard deviation.

$R^2$ is a correlation measure representing how well the model fits the data. RMSE is a measure of the overall error of the model. Consideration of the bias and σ enables one to further

decompose the error. The bias is an estimate of the systematic error, whilst σ is an estimate of the random error of the model.

In our analysis we will consider a statistically useful prediction to be one in which the RMSE of the prediction is lower than the standard deviation of the experimental data. If this criterion is not met, then the computational predictions are less accurate than the null model in which all molecules are predicted to have the same value as the mean of the experimental data. Whilst this is a useful statistical definition that provides a lower bound by which to define a useful model, it does not necessarily imply that methods that perform better than the null model are completely satisfactory.

## Results and Discussion

This section is set out in the following manner: Firstly QSPR predictions are presented using only molecular descriptors for the SUB-158 data. Unfortunately we cannot carry out theoretical chemistry predictions for SUB-158 due to a lack of suitable crystallographic input structures. These QSPR predictions, over the SUB-158 dataset, provide a useful frame of reference in which to consider the accuracy of QSPR predictions for the different sublimation terms (enthalpy, entropy and free energy). Secondly, we provide predictions for the SUB-48 dataset. We can perform theoretical chemistry calculations for the molecules in this dataset. We present predictions exclusively from the theoretical chemistry methods, followed by QSPR model predictions using one of three descriptor sets in turn: theoretical chemistry terms, CDK descriptors and both theoretical chemistry terms and CDK descriptors. We note at this point that the dataset is small and that ideally this would be done over a larger dataset. This analysis demonstrates the use of a different source of descriptors i.e. theoretical chemistry. Due to the size of this dataset the statistical power is low, but the results provide an indication of what may be expected from the incorporation of theoretical chemistry terms

as descriptors for this problem. A comparison and analysis of the SUB-48 and SUB-158 results is then given. Finally, we provide a summary discussion of the errors found in these models with reference to the Supporting Information.

*SUB – 158 Dataset QSPR Predictions*

**Tables 1 - 3** and **Figures 2**, **4** and **6** present the QSPR predictions for enthalpy, entropy and free energy of sublimation over the SUB-158 dataset. For each prediction a plot of the best performing model is given. The Supporting Information contains plots for other models. **Figures 3**, **5** and **7** show a summary boxplot of the ten most important descriptors found in the RF algorithm for each predicted property. The variable importance was calculated in the R package randomForest. [107]

**Table 1**. Enthalpy of sublimation predictions using QSPR models for the SUB-158 dataset (RMSE in kJ/mol), σ is the standard deviation. Experimental σ = 17.81 kJ/mol, range = 105.80 kJ/mol.

| Data/Measure | RF ± σ | SVM ± σ | PLS ± σ |
|---|---|---|---|
| CDK $R^2$ | 0.62 ± 0.01 | 0.56 ± 0.02 | 0.65 ± 0.02 |
| CDK RMSE | 11.19 ± 0.15 | 11.95 ± 0.33 | 10.71 ± 0.33 |

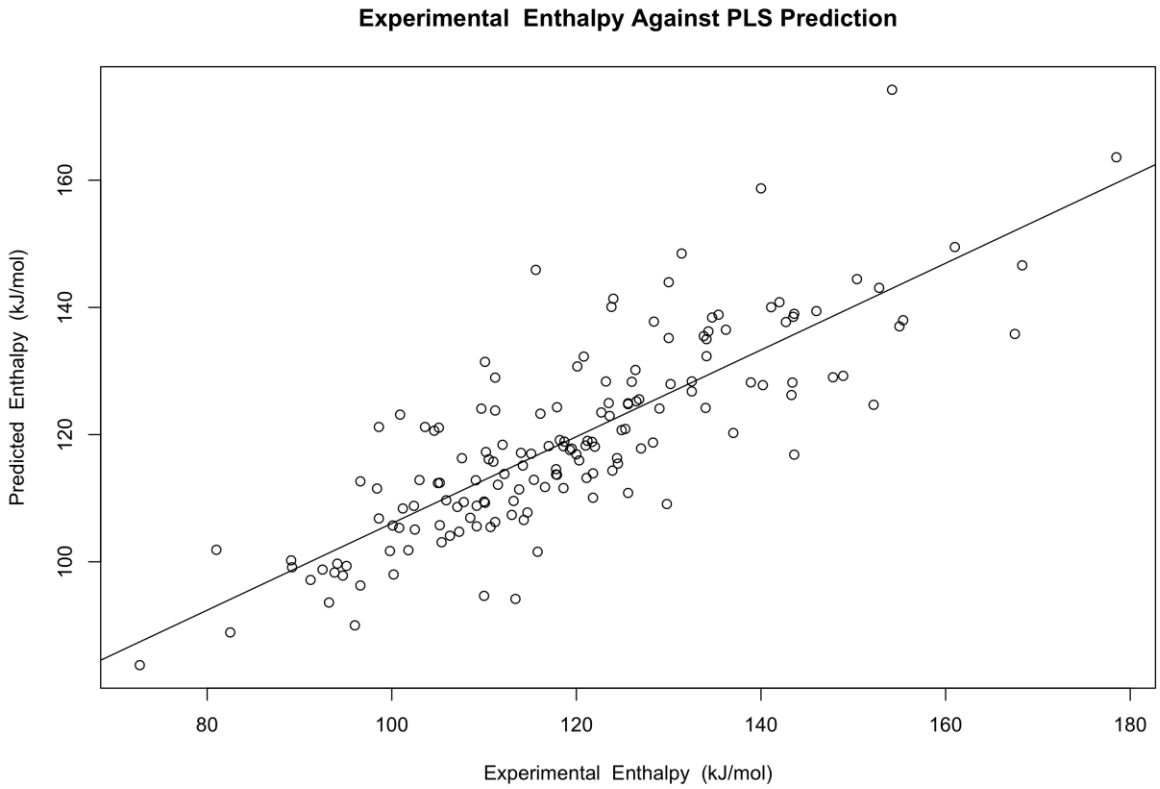



**Figure 2.** Enthalpy of sublimation predictions using the PLS model for the SUB-158 dataset. (see **Figures S1 – S3** for plots of all models)

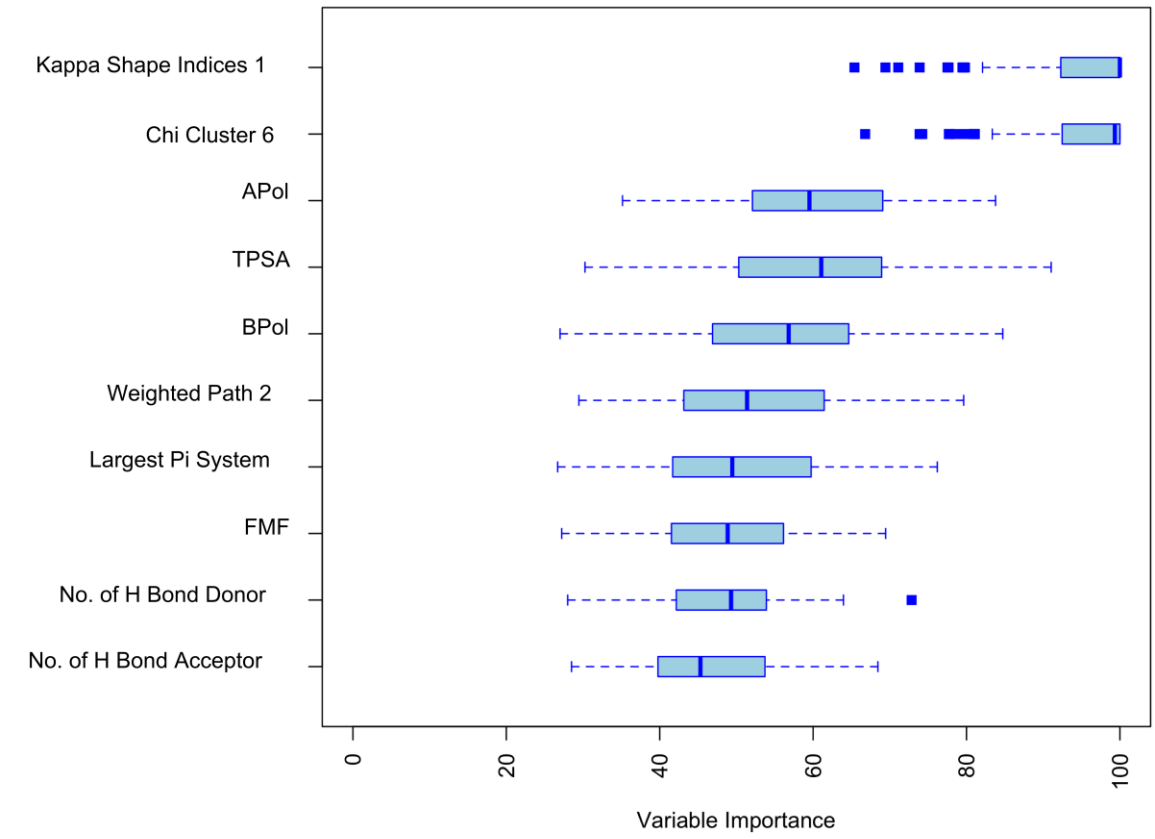



**Figure 3.** Enthalpy of sublimation variable importance from RF. The x-axis displays the percentage of predictions in which a descriptor is rated as important with the box and whiskers representing the 95$^{th}$ percentile. Dark blue lines in the boxes represent the median and the blue boxes extreme values. The y-axis states the descriptor.

Above, the PLS model provides the highest $R^2$ and lowest RMSE with σ values comparable to those of the other methods. We therefore suggest this is the best performing of the three models. The models all meet our statistical usefulness criterion, with RMSE's within the experimental standard deviation. We note several important topological descriptors such as Kappa shape indicia 1, Kier Hall cluster (SC.6),[51113] and the topological polar surface area (TPSA)[114] These descriptors make some physical sense describing the shape and polarity of a molecule.

**Table 2**. $T\Delta S_{sub}$ predictions using QSPR models for the SUB-158 dataset (RMSE in kJ/mol), σ is the standard deviation. Experimental σ = 10.79 kJ/mol, range = 63.13 kJ/mol.

| Data/Measure | RF ± σ | SVM ± σ | PLS ± σ |
|---|---|---|---|
| CDK $R^2$ | 0.62 ± 0.01 | 0.54 ± 0.03 | 0.48 ± 0.02 |
| CDK RMSE | 6.72 ± 0.08 | 7.37 ± 0.28 | 7.8 ± 0.13 |

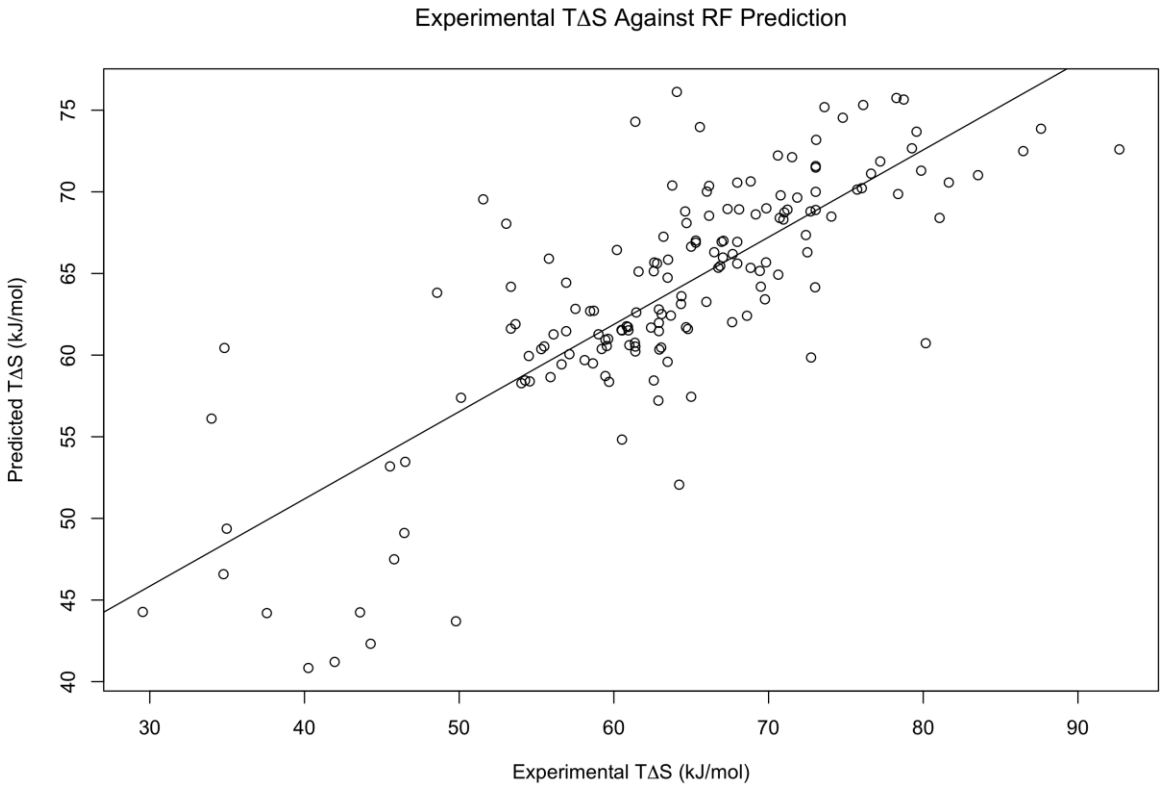



**Figure 4.** TΔS of sublimation predictions using the RF model for the SUB-158 dataset. (see **Figures S4 – S6** for plots of all models)

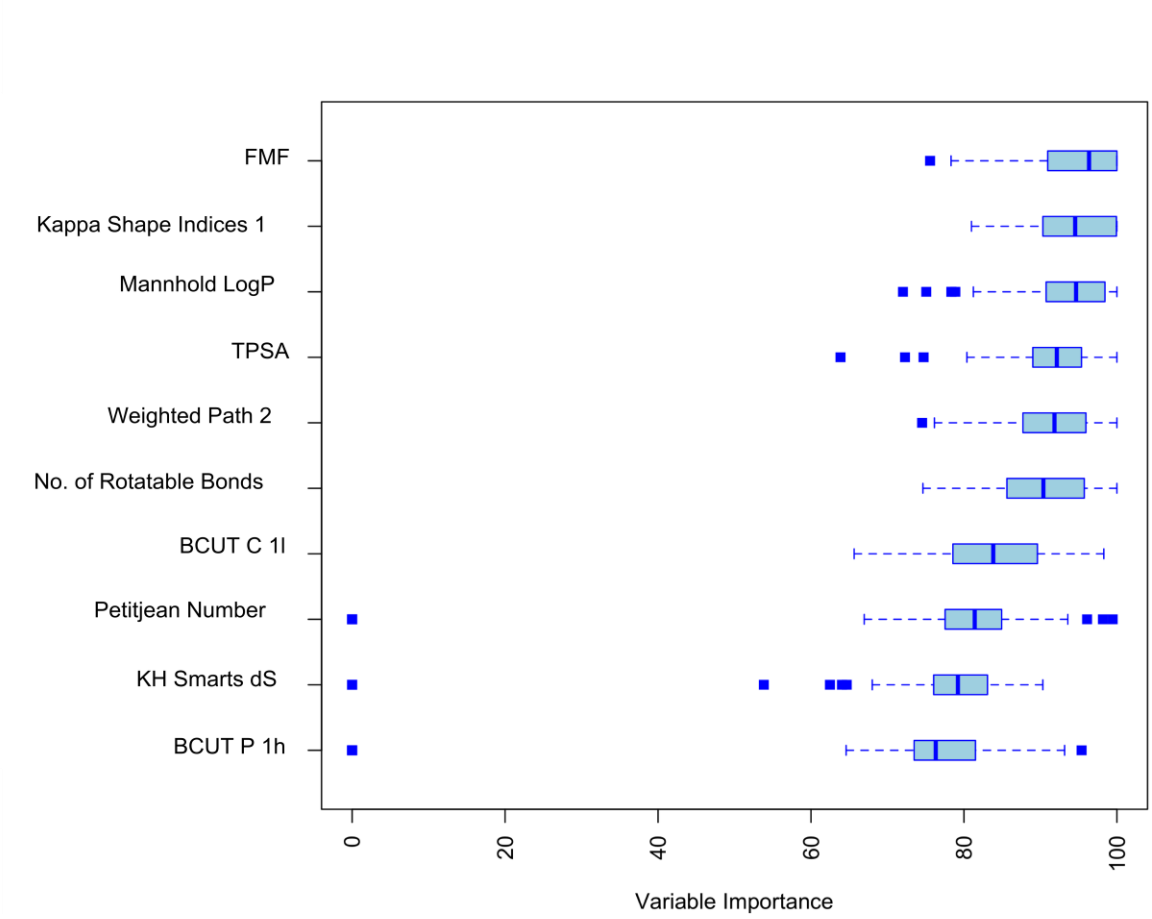



**Figure 5.** TΔS of sublimation variable importance from RF. The x-axis displays the percentage of predictions in which a descriptor is rated as important with the box and whiskers representing the 95$^{th}$ percentile. Dark blue lines in the boxes represent the median and the blue boxes extreme values. The y-axis states the descriptor.

Above, the RF model provides the highest $R^2$ and lowest RMSE with σ values equal or lower than σ values from the other two methods; hence we consider this model to be the best performing model. **Figure 5** expounds the variable importance from the RF model for the TΔS of sublimation. We find similar descriptors to be important to those for the enthalpy of sublimation, such as the Kier Hall kappa shape indices. Additionally, we find highly-ranked descriptors one might expect from a physical perspective such as the number of rotatable bonds (nRotB) and the weighted path (WTPT.4) descriptor. These descriptors relate directly to the flexibility and extent of a molecule. One may have expected to see the molecular weight also ranked highly, however this is not the case here.

**Table 3**. Free energy of sublimation predictions using QSPR models for the SUB-158 dataset (RMSE in kJ/mol). σ is the standard deviation. Experimental σ = 15.53 kJ/mol, range = 92.92 kJ/mol.

| Data/Measure | RF ± σ | SVM ± σ | PLS ± σ |
|---|---|---|---|
| CDK $R^2$ | 0.73 ± 0.01 | 0.64 ± 0.03 | 0.76 ± 0.02 |
| CDK RMSE | 8.21 ± 0.18 | 9.23 ± 0.42 | 7.55 ± 0.39 |

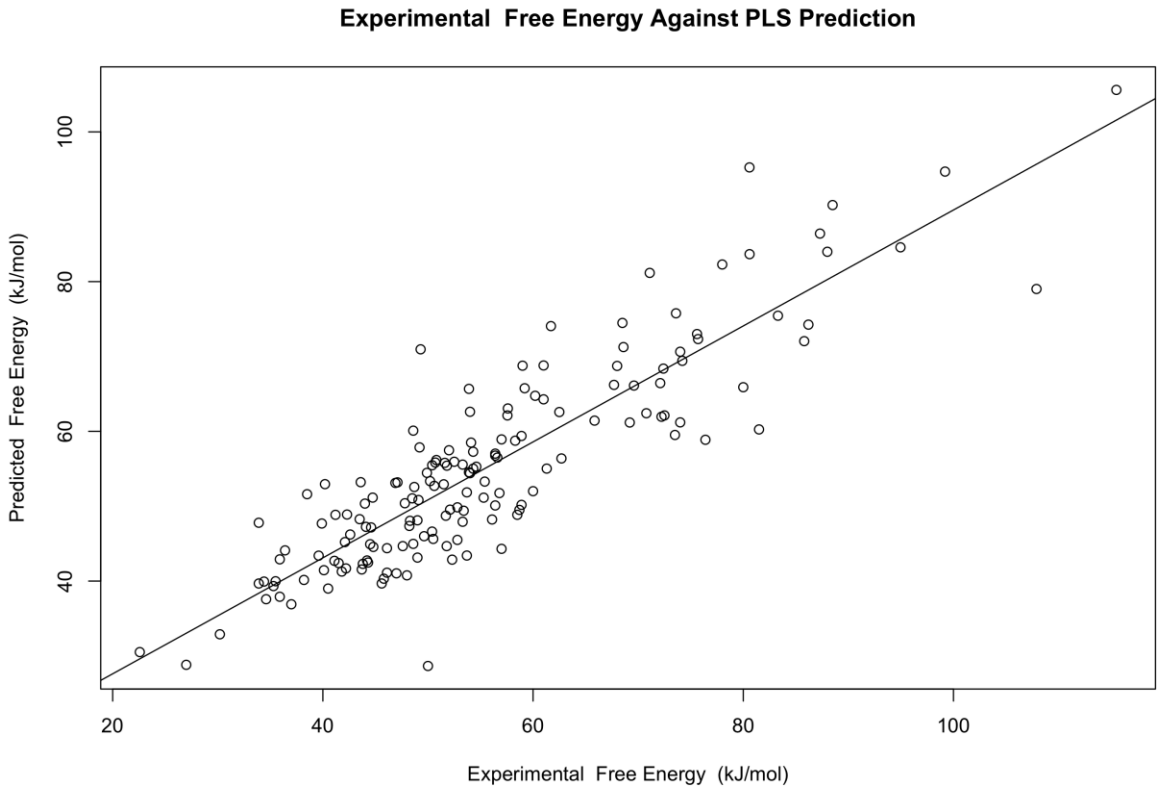



**Figure 6.** The free energy of sublimation predictions using the PLS model for the SUB-158 dataset. (see **Figures S7 – S9** for plots all models)

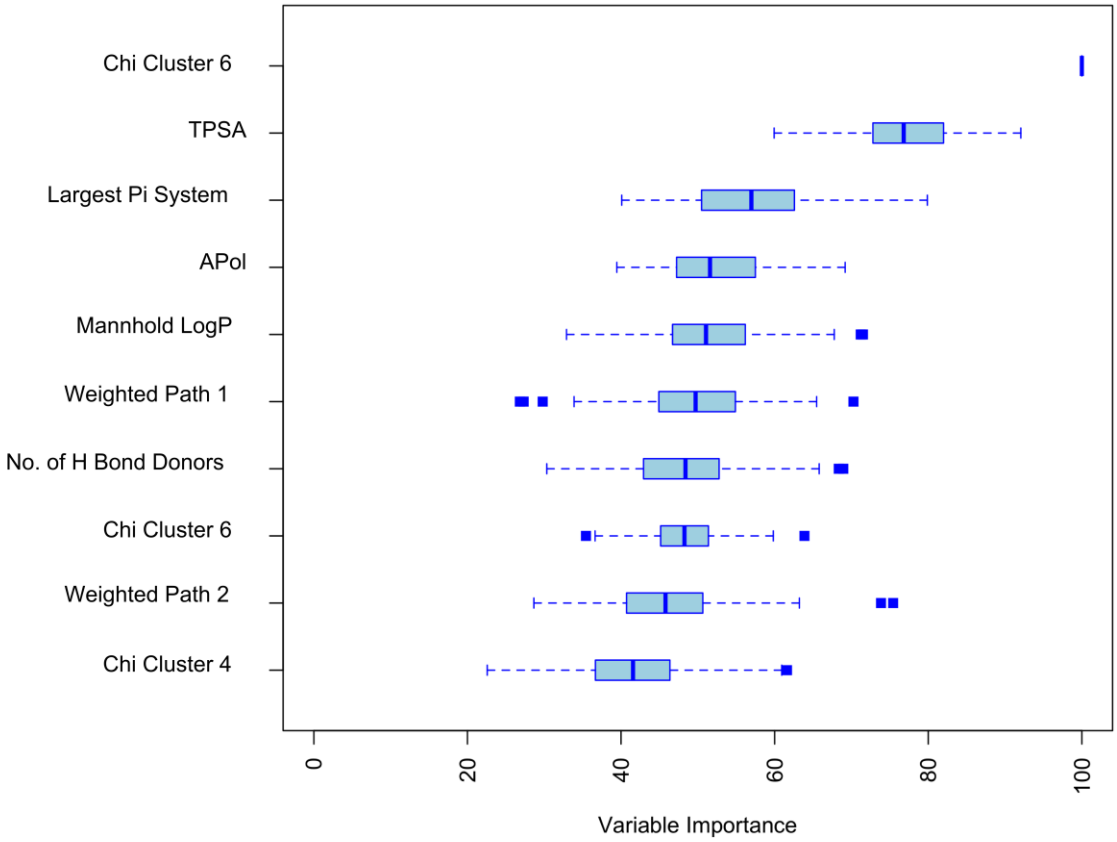



**Figure 7.** Free energy of sublimation variable importance from RF. The x-axis displays the percentage of predictions in which a descriptor is rated as important with the box and whiskers representing the 95$^{th}$ percentile. Dark blue lines in the boxes represent the median and the blue boxes extreme values. The y-axis states the descriptor.

The PLS model has superior $R^2$ and RMSE values compared to the RF and SVM models. We present the PLS model here as the best performing model. **Figure 7** summarizes the important descriptors for free energy of sublimation prediction using the RF algorithm on the SUB-158 dataset. We see that many of the descriptors important here were also important for the prediction of the enthalpy and/or entropy of sublimation.

We can see that in all cases the SUB-158 models make a useful prediction of the thermodynamics of sublimation, according to our statistically useful prediction criterion. We find all models making a useful prediction from CDK descriptors with an RMSE of approximately 11 kJ/mol. Whilst this is not a quantitatively useful level of accuracy, recent

work utilizing more advanced and computationally demanding descriptors, was able to achieve a standard error of prediction of 7.88 ± 0.35 kJ/mol.[8] We believe these results suggest the enthalpy of sublimation is a predictable property.

The entropy of sublimation appears the most difficult thermodynamic parameter to predict. Considering the differences between the RMSE of the best performing QSPR models and the experimental standard deviation, for each thermodynamic term over the SUB-158 dataset, we can see that proportionally the smallest difference between these terms is in the prediction of the TΔS values. This suggests the QSPR models are able to explain less of the variance in the TΔS data than either of the other properties. We find correlations comparable with those of the enthalpy predictions. PLS produces the poorest prediction, suggesting that linear models may not be suitable for predictions of the entropy of sublimation.

The free energy of sublimation is well predicted by PLS and RF. The predictions are all well inside the experimental standard deviation with good $R^2$ values. SVM makes a poorer prediction here than PLS and RF.

*Theoretical Chemistry Predictions of SUB-48*

**Figure 8** shows the prediction of the enthalpy of sublimation from theoretical chemistry calculations. The standard deviation of the experimental data is 15.94 kJ/mol, marginally greater than the RMSE from the predictions (15.26 kJ/mol). These results therefore just qualify as a statistically useful prediction, applying our criterion above. A reasonable positive correlation exists with $R^2$ = 0.56. Interestingly, σ is larger than the bias and therefore the largest contribution to the error is random error. The results lead to the conclusion that a large amount of random variation is found in the data, which cannot be explained by the present method.

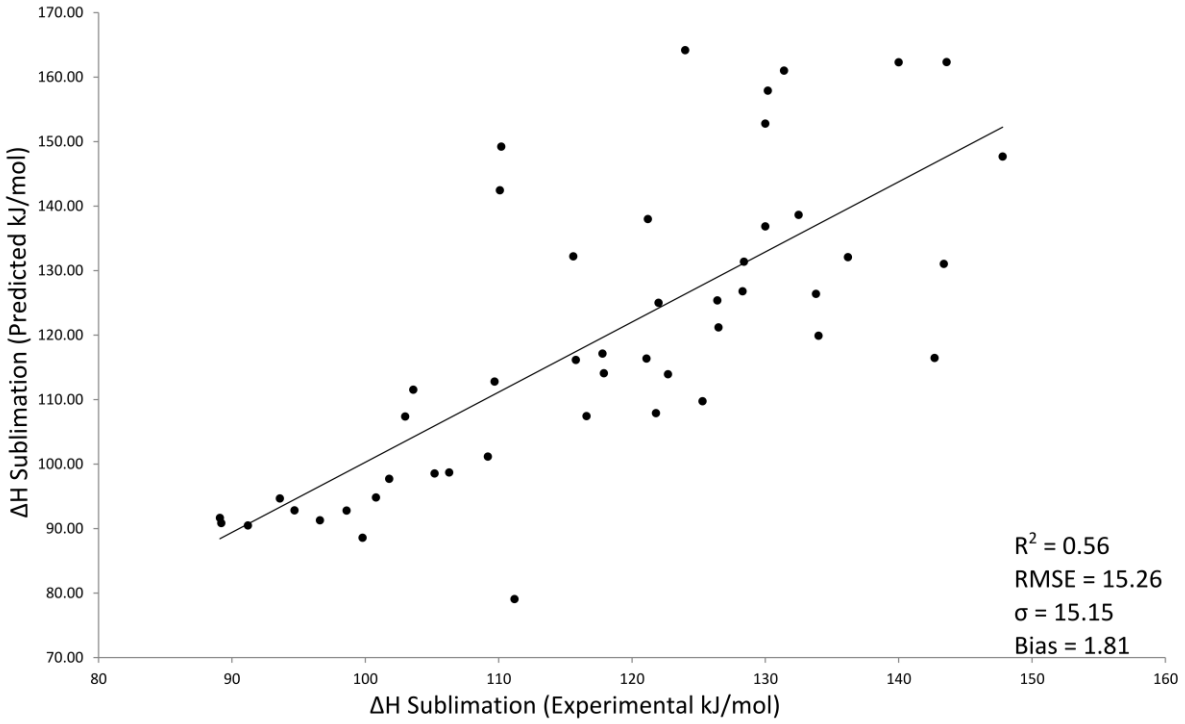



**Figure 8**. Predictions of $\Delta H_{sub}$ from theoretical chemistry. Experimental $\sigma$ = 15.94 kJ/mol

**Figure 8** shows the $T\Delta S_{sub}$ predictions made using theoretical chemistry methods. These predictions were made assuming rigid-body behavior, i.e. all intramolecular contributions are consistent over the phase transition.

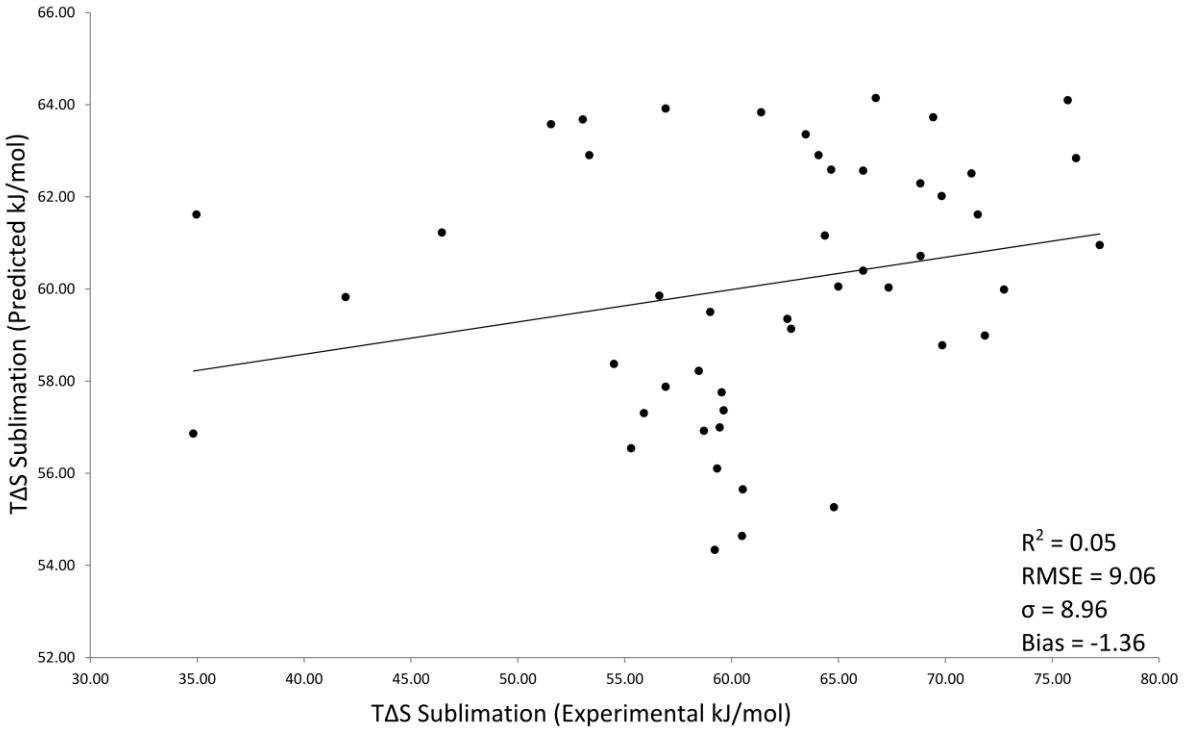



**Figure 9**. Predictions of $T\Delta S_{sub}$ from theoretical chemistry. Experimental σ = 9.33 kJ/mol.

Clearly, **Figure 9** shows no correlation between the predicted and experimental $T\Delta S_{sub}$. As with the enthalpy of sublimation, the RMSE is just marginally within the experimental data's σ (9.33 kJ/mol). Once again the major contribution to the RMSE comes from random errors. There are two possible causes for the poor results in **Figure 9**: First, there may be a vital contribution missing in our model, namely dynamic body as opposed to rigid body contributions; for instance, coupling between intra- and intermolecular modes. Second, the errors quoted in the experimental data may be too small. We note here that the experiments to obtain these values are very complex. The Supporting Information shows plots of enthalpy-entropy compensation (**Figure S10**) and molecular mass against $T\Delta S_{sub}$ (**Figure S11**). We applied constant corrections for the lack of internal motion in the crystalline models in the form of Rln(2) and Rln(3) terms per rotatable bond (**Figure S12**). This showed no real improvement in the prediction and so does not appear in the final models. We additionally

neglect energy associated with conformational change between phases. A global optimization in the vacuum could provide an improvement to these models.

Finally, we combine predictions from **Figures 8** and **9** to predict the free energy of sublimation.

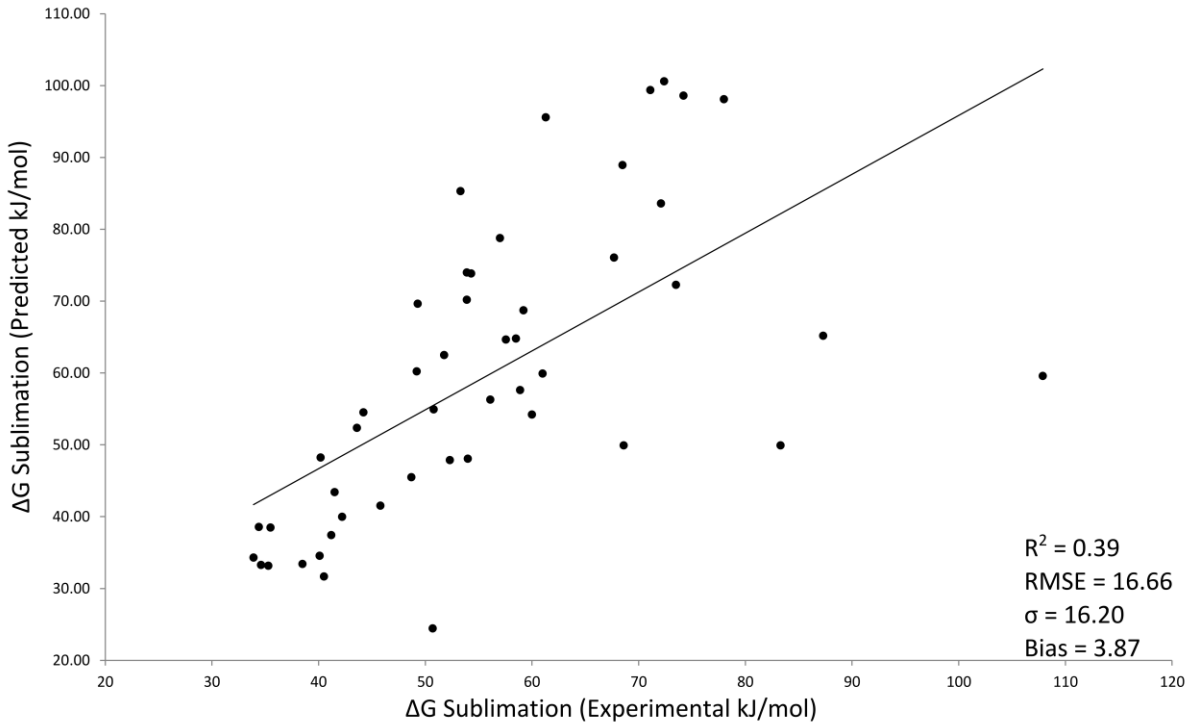



**Figure 10**. Predictions of $\Delta G_{sub}$ from theoretical chemistry. Experimental $\sigma$ = 15.61 kJ/mol.

**Figure 10** displays a weak correlation between the predictions and the experiment. As in **Figures 8** and **9,** the RMSE is largely due to random error. Additionally, the RMSE of the prediction is greater than the standard deviation of the experimental data, and fails to meet our criterion for a statistically useful prediction. The experimental standard deviation is 15.61 kJ/mol for the free energy of sublimation. We note that in all clearly outlying points, nitrogen containing functional groups are present (**Figure S13**), this is especially true of functional groups containing N=O moieties. In our experience these groups can be difficult to accurately

represent with the current model, which suggests a problem with the intermolecular potential for groups containing this atom.

Overall these predictions suggest it is not possible to quantitatively predict all of the sublimation thermodynamic properties using these approximate theoretical chemistry methods. Previous work has demonstrated the predictability of the enthalpy of sublimation[2,8,115] using QSPR methods on datasets varying in size from 213 to 1302,[8,115] which again is seen here to be the most predictable term, although with lower accuracy than previously reported. However, the entropy and ultimately the free energy of sublimation, appear to be poorly predicted. Predictions of $T\Delta S_{sub}$ show effectively no correlation with the experimental values. The free energy of sublimation, although missing our statistically useful prediction criterion, does show some correlation between predictions and experiment. It may be that employing more advanced theoretical chemistry methods, such as periodic DFT, could improve this situation, although at a much increased computational cost (second derivatives will be required for phonon mode calculations). Compared to the QSPR predictions over the SUB-158 dataset, these predictions are far inferior, suggesting a more advance quantum mechanical method needs to be employed to achieve useful predictions from theoretical chemistry alone. It is not clear from these results if the major error is a result of noisy or imprecise data, or emanates from the approximations and methodological options employed in the modeling. The experimental and predicted results are all reported in the Supporting Information (**Table S4**).

*QSPR Predictions of SUB-48*

**Tables 4 – 6** present the results of the QSPR models developed using the SUB-48 dataset. We present 27 QSPR models (three machine learning algorithms tested on three input datasets for each of three thermodynamic quantities: enthalpy in **Table 4**, $T\Delta S$ in **Table 5** and

free energy in **Table** 6 3x3x3=27 models). An inter-descriptor correlation cut-off of 0.9 was applied all datasets. The remaining features under went auto-scaling (standardization of the mean and standard deviation of each descriptor) on all datasets. The different datasets are represented in the first column of the tables: Predicted Thermodynamics (PT) refers to the dataset that uses thermodynamic values predicted from the previous section using theoretical chemistry methods as descriptors. Chemistry Development Kit (CDK) refers to the dataset that uses CDK descriptors, calculated from a SMILES representation of the molecules in the dataset. All refers to both sets of descriptors used together.

**Table 4**. Enthalpy of sublimation predictions using QSPR models for SUB-48 (RMSE kJ/mol). σ is the standard deviation. Experimental σ = 15.94 kJ/mol, range = 58.70 kJ/mol. (**Figures S14–S22**)

| Dataset/Measure | RF (± σ) | SVM (± σ) | PLS (± σ) |
|---|---|---|---|
| PT $R^2$ | 0.53 ± 0.03 | 0.49 ± 0.04 | 0.44 ± 0.04 |
| PT RMSE | 10.92 ± 0.35 | 11.27 ± 0.48 | 11.88 ± 0.45 |
| CDK $R^2$ | 0.37 ± 0.03 | 0.44 ± 0.04 | 0.33 ± 0.06 |
| CDK RMSE | 12.46 ± 0.31 | 11.86 ± 0.57 | 13.35 ± 0.9 |
| All $R^2$ | 0.56 ± 0.03 | 0.54 ± 0.02 | 0.36 ± 0.06 |
| All RMSE | 10.45 ± 0.31 | 10.69 ± 0.29 | 13.08 ± 0.81 |

**Table 5**. $T\Delta S_{sub}$ predictions using QSPR models for SUB-48 (RMSE kJ/mol). σ is the standard deviation. Experimental σ = 9.33kJ/mol, range = 42.40kJ/mol. (**Figures S23–S31**)

| Dataset/Measure | RF (± σ) | SVM (± σ) | PLS (± σ) |
|---|---|---|---|
| PT $R^2$ | 0.01 ± 0.01 | 0.01 ± 0.01 | 0.01 ± 0.02 |
| PT RMSE | 9.71 ± 0.34 | 9.82 ± 0.6 | 9.63 ± 0.26 |
| CDK $R^2$ | 0.29 ± 0.06 | 0.05 ± 0.03 | 0.26 ± 0.05 |
| CDK RMSE | 7.77 ± 0.28 | 9.17 ± 0.29 | 8.01 ± 0.31 |
| All $R^2$ | 0.32 ± 0.07 | 0.06 ± 0.04 | 0.26 ± 0.06 |
| All RMSE | 7.62 ± 0.35 | 9.06 ± 0.31 | 7.97 ± 0.33 |

**Table 6**. Free energy of sublimation predictions using QSPR models for SUB-48 (RMSE kJ/mol). σ is the standard deviation. Experimental σ = 15.61 kJ/mol, range = 74.00 kJ/mol. (**Figures S32–S40**)

| Dataset/Measure | RF (± σ) | SVM (± σ) | PLS (± σ) |
|---|---|---|---|
| PT $R^2$ | 0.27 ± 0.02 | 0.31 ± 0.02 | 0.16 ± 0.04 |
| PT RMSE | 13.59 ± 0.33 | 13.04 ± 0.29 | 14.49 ± 0.63 |
| CDK $R^2$ | 0.48 ± 0.03 | 0.37 ± 0.05 | 0.5 ± 0.03 |
| CDK RMSE | 11.13 ± 0.25 | 12.34 ± 0.55 | 10.97 ± 0.4 |
| All $R^2$ | 0.57 ± 0.02 | 0.47 ± 0.04 | 0.5 ± 0.08 |
| All RMSE | 10.17 ± 0.17 | 11.26 ± 0.38 | 11.05 ± 1.08 |

**Table 6** shows consistent improvement in predictive accuracy related to the free energy of sublimation when compared to the theoretical chemistry methods alone. In all cases the statistical usefulness prediction criterion is now met, i.e. for all predictions the RMSE is lower than the experimental standard deviation (15.61 kJ/mol). It has previously been reported that for solubility, combining descriptors from theoretical chemistry with 2D cheminformatics descriptors does not notably improve the model (the descriptor sets were non-complementary).[10] However, for sublimation thermodynamics, it seems that the two descriptor sets are complementary. A reduction in prediction RMSE is accompanied by an improvement in $R^2$ for the models that combine the two descriptor sets. This suggests that the incorporation of descriptors specifically related to the crystal can have an impact on the predictive accuracy. Despite this improvement, the overall level of predictive accuracy is still fairly low for most applications, with sizeable RMSE values found even after coupling the two descriptor sets. We therefore suggest that whilst these PT descriptors fail to provide the

level of improvement required, these preliminary results suggest that there is a potential complementarity between such descriptor sets.

In **Tables 4** and **5** we see a divergence in the predictability between the enthalpy of sublimation and $T\Delta S_{sub}$ respectively. All predicted enthalpy of sublimation RMSE values are well within the experimental standard deviation. This result is in agreement with previous work suggesting that the enthalpy of sublimation is predictable by QSPR methods.[8,116] It seems that, in terms of enthalpy of sublimation predictions, the 2D CDK descriptors provide either similar information in the case of SVM or less information in the case of RF to the models compared to the PT descriptors, as only a modest improvement is found on combining the descriptor sets. When the PLS algorithm is employed the prediction becomes worse when 2D CDK descriptors are used, suggesting a level of redundancy in the descriptor set. PLS consistently makes a worse prediction of the enthalpy of sublimation compared to the other methods. This was not seen in the SUB-158 dataset, suggesting it is an artefact of the small dataset.

The opposite appears true when we consider $T\Delta S_{sub}$. In this case, 2D CDK descriptors appear to offer more information to the machine learning models than the PT descriptors. The model built on the PT descriptors actually shows zero correlation and represents a useless model in terms of predictability. Perhaps this should not be such a surprise given the poor correlation between the theoretical chemistry predictions and experimental $T\Delta S_{sub}$. All algorithms fail to provide a reasonable prediction with only the PT descriptors. The RMSE values for all machine learning algorithms dramatically improve when provided with 2D CDK descriptors, with SVM showing the least improvement. All models provide a statistically useful prediction of $T\Delta S_{sub}$, according to our criterion, when the descriptor sets are combined (All). However, the low $R^2$ values show a poor correlation at best. We believe that, as a reasonable prediction can be made by these algorithms with the SUB-158 dataset, the literature values do

correspond to genuine entropic contributions; however, it is possible that the theoretical chemistry terms used here express only partial contributions to the experimental value of $T\Delta S_{sub}$. RF and PLS both outperform SVM in this task. None of these machine learning algorithms produce a statistically useful prediction, judged by our criterion, of $T\Delta S_{sub}$ from the PT descriptors alone. Whilst we acknowledge the poor quality of these predictions, hence the descriptor importance cannot be assumed to be generally important to $T\Delta S_{sub}$ predictions, **Figures S41** and **S42** present the descriptors rated as important in this case. Several descriptors were consistently rated amongst the most important in $T\Delta S_{sub}$ predictions: weighted path, the number of nitrogen atoms, Kier Hall smarts (group counting based on molecular fragmentation) and topological surface area (TPSA). When PT descriptors are provided in the all descriptor set $\Delta S^{PT}$ and $\Delta G^{PT}$ also feature in the top ten most important variables. From a physical standpoint, one might expect to find descriptors such as the number of rotatable bonds and molecular mass rated as highly important, as molecules with a high number for either of these properties would generally be expected to have larger entropy. Whilst the number of rotatable bonds does feature in the top ten important descriptors when CDK descriptors are used alone, the molecular weight does not. However, weighted path descriptors, which describe the degree of molecular branching, do feature as the top descriptor. These descriptors differ from those found to be important for the larger SUB-158 dataset above in several cases. We note also occurrences in **Table 5** in which the standard deviation of the $R^2$ values implies a potentially negative $R^2$ value. This is an artefact of the calculation suggesting there is no correlation i.e. $R^2$ is zero, hence only the upper bound should be considered as a valid value. It is possible, given the small dataset resulting from the limited amount of information available fulfilling our criteria, that in some of these cases the models have been over fitted.

These results provide evidence for the predictability of the enthalpy of sublimation. Previous work has shown this quantity to be accessible via QSPR models[1,2,7,116] to reasonable accuracy. In the current work we show that the addition of theoretical chemistry terms as descriptors provides an improvement in the prediction, above that of the conventional 2D descriptors from the CDK. We therefore suggest values from theoretical chemistry could be considered in the future as descriptors for QSPR models of sublimation thermodynamics (enthalpy and free energy of sublimation) where experimental (or predicted) data permit. The computational cost of the theoretical chemistry procedure (a few hours per molecule) is expensive compared to some descriptor calculations, but with an improvement of up to 16.1% (RF(all):RF(CDK) enthalpy SUB-48) and 8.6% improvement (RF(all):RF(CDK) free energy SUB-48) in the RMSE, we feel this computational time is worthwhile. $T\Delta S_{sub}$ shows mixed behavior. This quantity is only poorly predicted by RF and PLS, but owes much more of the useful information to the CDK descriptors rather than the theoretical calculations. It appears that the theoretical quantities used here as alternative descriptors do not provide much, if any, useful information above that of the descriptors from the CDK. However, given the poor correlation of the theoretical predictions with experiment this is not so surprising and suggests that a more sophisticated first principles theory is required to provide entropy values of greater accuracy and use. However, these more sophisticated first principles methods can be extremely expensive and therefore are not amenable to use in QSPR models. Alternative descriptors are required to capture information about the entropic contributions.

In summary, it appears that the enthalpy and free energy can be modeled using QSPR methods to satisfy our statistical usefulness criterion with CDK descriptors alone (see SUB-158). For these terms the SUB-48 results show a potential complementarity between the descriptor sets of CDK descriptors and PT descriptors specific to the crystal structure. The level of predictive accuracy over the SUB-48 data however, falls below that required for

quantitative use and the small sample size prevents us from suggesting this as a general finding. The entropy of sublimation appears to be less predictable. It is possible that a more rigorous calculation may be needed to achieve an accurate prediction; there may also be significantly larger errors in the experimental data than are reported in the literature. Methods capable of such predictions are likely to be too expensive for use in QSPR models.

Compared to the SUB-158 predictions, the SUB-48 predictions are poor. Indeed for the entropy of sublimation the models based upon PT descriptors are in effect useless, which is unsurprising given the poor correlation between the PT values and the experimental values for $T\Delta S_{sub}$. However, these preliminary results have suggested a potential complementarity between the CDK descriptors and the PT descriptors, indicating that further exploration of QSPR modeling for free energy and enthalpy of sublimation will be worthwhile.

*Analysis of the SUB-48 TΔS predictions*

To help to determine the sources of the errors in the entropy predictions and data, we have plotted the experimental and predicted TΔS against physical properties one may expect it to correlate with, namely number of rotatable bonds and molecular mass. Additionally we provide a QSPR prediction of the $T\Delta S_{sub}$ values predicted using theoretical chemistry methods, using the same molecular descriptors as the results presented above (i.e. replace the experimental data with the DMACRYS-G09 predictions to see to what extent CDK descriptors can model these predictions).

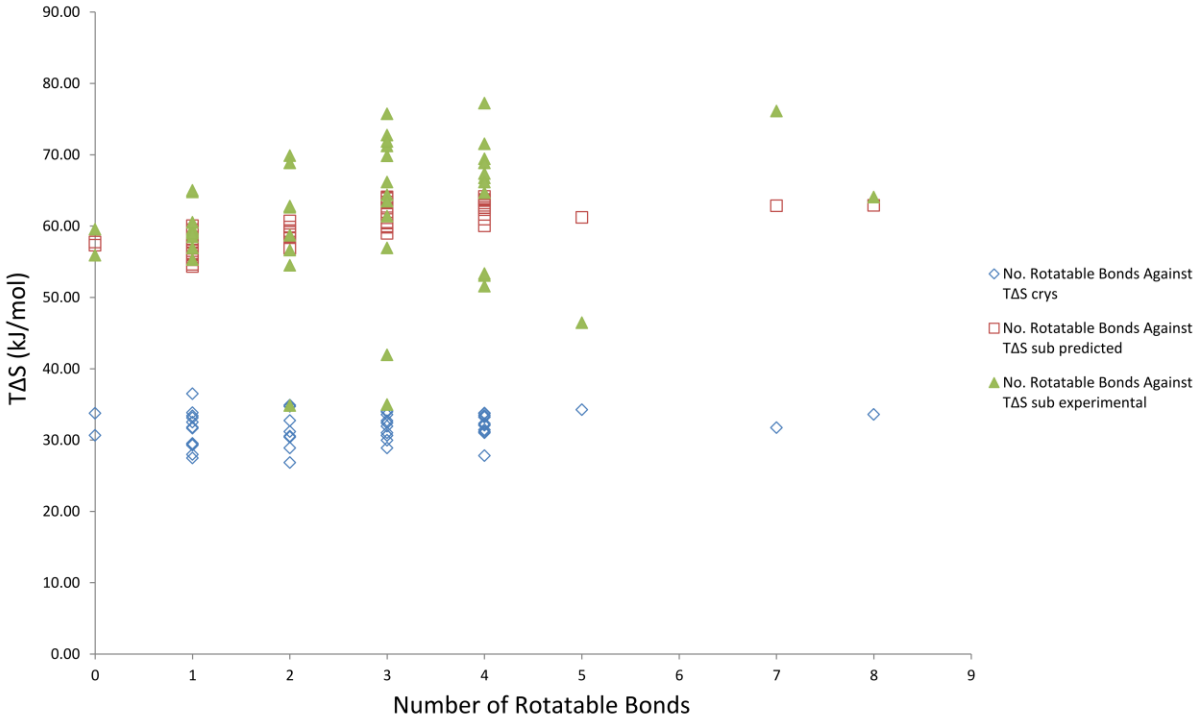



**Figure 11**. The number of rotatable bonds against TΔS of sublimation.

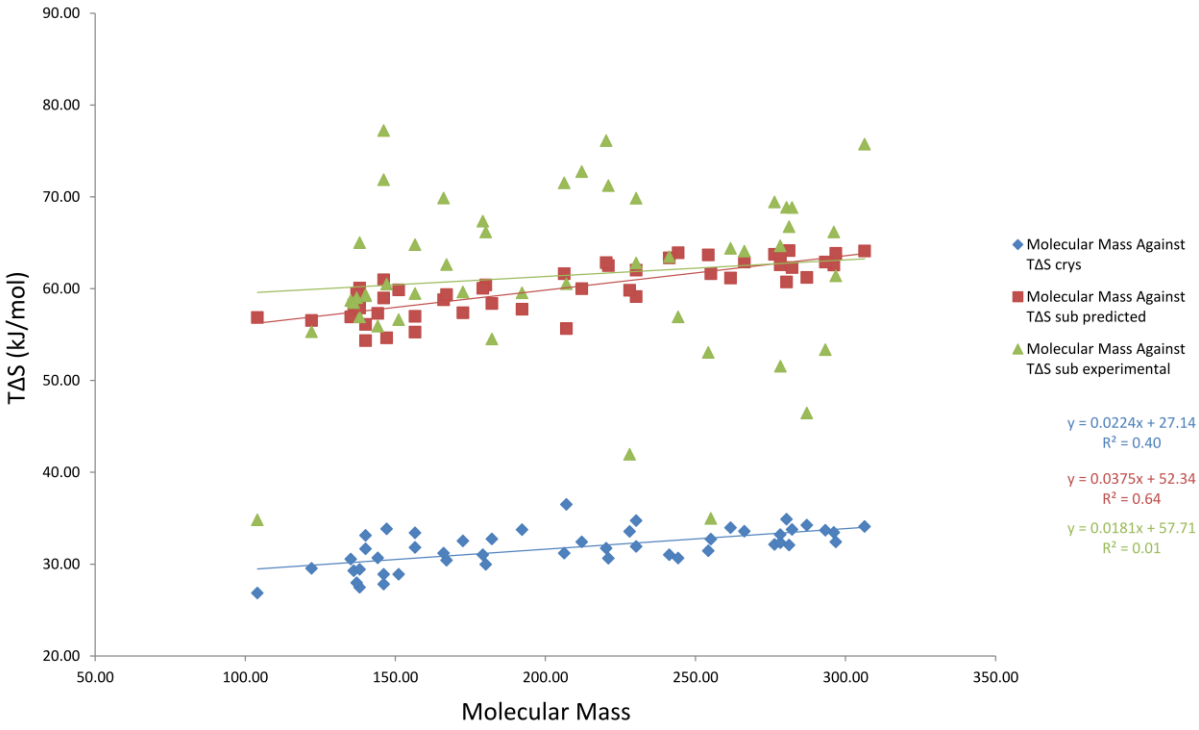



**Figure 12**. Molecular mass against TΔS of sublimation.

**Table 7.** QSPR predictions of calculated DMACRYS-G09 TΔS RMSE presented in kJ/mol. DMACRYS-G09 σ = 2.82 kJ/mol, range = 9.76 kJ/mol.

| Dataset/Measure | RF (± σ) | SVM (± σ) | PLS (± σ) |
|---|---|---|---|
| CDK $R^2$ | 0.78 ± 0.02 | 0.67 ± 0.03 | 0.73 ± 0.02 |
| CDK RMSE | 1.31 ± 0.06 | 1.61 ± 0.06 | 1.46 ± 0.05 |

We can see in both **Figure 11** and **12** that the predicted entropy values cluster around quite a tight region of the TΔS axis spanning ~10 kJ/mol. The experimental TΔS values vary much more dramatically, spanning a range of over 40 kJ/mol. Given that one would naturally expect a reasonable correlation between TΔS and the molecular mass or number of rotatable bonds, it seems odd that the experimental terms do not reflect this. The average error margin for the TΔS of sublimation quoted in the original literature source for the SUB-48[52–105] dataset is 0.79 kJ/mol (range 0.09 –2.68 kJ/mol). The full SUB-48 data set and references to the original sources are given in the Supporting Information **Table S1**. These data are not enough to conclude that the difficulty in the prediction of entropy is due to noise from different experimental techniques. Our own prediction methods are inherently approximate and a perfect correlation is therefore not expected. Our own methods also contain a number of assumptions which are unlikely to be generalizable across all of chemical space, but which are accepted as useable assumptions within the community, such as rigid body behavior. It is highly likely that this assumption has a notable effect on the accuracy of the TΔS prediction. **Table 7** shows that there is a good correlation with approximately 1.5 kJ/mol RMSE's in the QSPR predictions of the calculated $T\Delta S_{sub}$ from theoretical chemistry methods. This is a much improved prediction compared to the equivalent predictions of the experimental data. The predictions here easily meet our statistically useful prediction criterion. This may well be expected given the large reduction in range of the $T\Delta S_{sub}$ values calculated from theoretical chemistry compared to experiment. These data do suggest it would be useful to have a

standard dataset provided by experimental colleagues in order to test prediction models in the future.

*Y-scrambling of the SUB-158 and SUB-48 datasets*

Table 8 presents the results of a y-scrambled run for all of the models. The predictions presented above are summarised next to the y-scrambled results for comparison purposes.

**Table 8.** Prediction and y-scrambling data for all models.

| | | **Predictions** | | | **Y-scambled** | | |
|---|---|---|---|---|---|---|---|
| **Dataset** | **Measure** | **RF ± σ** | **SVM ± σ** | **PLS ± σ** | **RF ± σ** | **SVM ± σ** | **PLS ± σ** |
| **SUB-158 enthalpy autoscale** | CDK R2 | 0.62 ± 0.01 | 0.56 ± 0.02 | 0.65 ± 0.02 | 0.01 ± 0.01 | 0 ± 0 | 0 ± 0 |
| | CDK RMSE | 11.19 ± 0.15 | 11.95 ± 0.33 | 10.71 ± 0.33 | 18.28 ± 0.17 | 18.27 ± 0.2 | 18.57 ± 0.28 |
| **SUB-158 entropy autoscale** | CDK R2 | 0.62 ± 0.01 | 0.54 ± 0.03 | 0.48 ± 0.02 | 0.02 ± 0.01 | 0.01 ± 0.01 | 0.03 ± 0.02 |
| | CDK RMSE | 6.72 ± 0.08 | 7.37 ± 0.28 | 7.80 ± 0.13 | 11.55 ± 0.13 | 11.01 ± 0.07 | 11.64 ± 0.18 |
| **SUB-158 free energy autoscale** | CDK R2 | 0.73 ± 0.01 | 0.64 ± 0.03 | 0.76 ± 0.02 | 0.02 ± 0.01 | 0.01 ± 0.01 | 0 ± 0.01 |
| | CDK RMSE | 8.21 ± 0.18 | 9.23 ± 0.42 | 7.55 ± 0.39 | 15.63 ± 0.18 | 15.75 ± 0.24 | 16.92 ± 0.25 |
| **SUB-48 enthalpy all descriptors autoscale** | All R2 | 0.56 ± 0.03 | 0.54 ± 0.02 | 0.36 ± 0.06 | 0.04 ± 0.02 | 0.03 ± 0.02 | 0 ± 0.01 |
| | All RMSE | 10.45 ± 0.31 | 10.69 ± 0.29 | 13.08 ± 0.81 | 17.45 ± 0.33 | 17.09 ± 0.55 | 17.17 ± 0.3 |
| **SUB-48 enthalpy CDK descriptors autoscale** | CDK R2 | 0.37 ± 0.03 | 0.44 ± 0.04 | 0.33 ± 0.06 | 0.01 ± 0.01 | 0.04 ± 0.03 | 0.03 ± 0.03 |
| | CDK RMSE | 12.46 ± 0.31 | 11.86 ± 0.57 | 13.35 ± 0.9 | 15.95 ± 0.22 | 17.47 ± 0.74 | 16.79 ± 1.53 |
| **SUB-48 enthalpy PT descriptors autoscale** | PT R2 | 0.53 ± 0.03 | 0.49 ± 0.04 | 0.44 ± 0.04 | 0 ± 0 | 0.01 ± 0.02 | 0.01 ± 0.01 |
| | PT RMSE | 10.92 ± 0.35 | 11.27 ± 0.48 | 11.88 ± 0.45 | 17.29 ± 0.36 | 16 ± 0.38 | 16.48 ± 0.51 |
| **SUB-48 entropy all descriptors autoscale** | All R2 | 0.32 ± 0.07 | 0.06 ± 0.04 | 0.26 ± 0.06 | 0.02 ± 0.02 | 0.04 ± 0.03 | 0.04 ± 0.03 |
| | All RMSE | 7.62 ± 0.35 | 9.06 ± 0.31 | 7.97 ± 0.33 | 9.32 ± 0.27 | 9.91 ± 0.28 | 9.68 ± 0.49 |
| **SUB-48 entropy CDK** | CDK R2 | 0.29 ± 0.06 | 0.05 ± 0.03 | 0.26 ± 0.05 | 0.04 ± 0.02 | 0.04 ± 0.02 | 0.01 ± 0.01 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **descriptors autoscale** | CDK RMSE | 7.77 ± 0.28 | 9.17 ± 0.29 | 8.01 ± 0.31 | 9.17 ± 0.24 | 9.11 ± 0.17 | 9.94 ± 0.43 |
| **SUB-48 entropy PT descriptors autoscale** | PT R2 | 0.01 ± 0.01 | 0.01 ± 0.01 | 0.01 ± 0.02 | 0.01 ± 0.01 | 0.02 ± 0.02 | 0.02 ± 0.02 |
| | PT RMSE | 9.71 ± 0.34 | 9.82 ± 0.6 | 9.63 ± 0.26 | 9.78 ± 0.29 | 9.24 ± 0.13 | 9.62 ± 0.2 |
| **SUB-48 free energy all descriptors autoscale** | All R2 | 0.57 ± 0.02 | 0.47 ± 0.04 | 0.5 ± 0.08 | 0.01 ± 0.03 | 0.02 ± 0.03 | 0.01 ± 0.01 |
| | All RMSE | 10.17 ± 0.17 | 11.26 ± 0.38 | 11.05 ± 1.08 | 16.29 ± 0.66 | 16.26 ± 0.26 | 17.7 ± 0.7 |
| **SUB-48 free energy CDK descriptors autoscale** | CDK R2 | 0.48 ± 0.03 | 0.37 ± 0.05 | 0.5 ± 0.03 | 0.01 ± 0.01 | 0 ± 0 | 0.04 ± 0.01 |
| | CDK RMSE | 11.13 ± 0.25 | 12.34 ± 0.55 | 10.97 ± 0.4 | 15.94 ± 0.36 | 16.4 ± 0.47 | 16.23 ± 0.39 |
| **SUB-48 free energy PT descriptors autoscale** | PT R2 | 0.27 ± 0.02 | 0.31 ± 0.02 | 0.16 ± 0.04 | 0.02 ± 0.02 | 0 ± 0 | 0 ± 0 |
| | PT RMSE | 13.59 ± 0.33 | 13.04 ± 0.29 | 14.49 ± 0.63 | 16.16 ± 0.43 | 16.26 ± 0.34 | 16.01 ± 0.2 |

y-scrambling was performed in order to test the results of the predictions compared against randomised prediction targets. One would expect to see a notable reduction in $R^2$ and an increase in the RMSE. Looking at the raw numbers, one does indeed find a notable change in the expected directions for all cases with the exception of the $T\Delta S_{sub}$ predicted by the PT descriptors, in which the results remain similar making clear that the original model was very poor.

*Discussion and comparison of SUB-158 and SUB-48*

In terms of models available in the literature, which have explored related methodologies, we believe that the models presented here offer a new insight. Recent work has presented excellent QSPR models for the prediction of sublimation enthalpy. Those models utilize a multiple linear regression methodology and four descriptor variables. That model demonstrates a much more accurate QSPR prediction than the models for the enthalpy of sublimation in this work. However, as was seen in previous work we also find a fairly linear

relationship between our descriptors and the property when we consider the statistical results showing PLS performing well on the SUB-158 dataset. We also find a generally positive correlation between the number of rotatable bonds and the thermodynamic terms (**Figures S43 – S45).** Our work additionally tackles the prediction of TΔS sublimation and the free energy of sublimation. These terms have received less attention than the enthalpy of sublimation.

Perlovich *et al.*[2] provided QSPR models for the enthalpy and free energy of sublimation utilizing the HYBOT descriptors. The entropy of sublimation was then estimated as the difference between the enthalpy and free energy. This work also generally finds an improved prediction of the free energy compared to the enthalpy of sublimation.

Salahinejad *et al.*[7] created a very promising model for intrinsic aqueous solubility prediction over a diverse chemical space and investigated the incorporation of lattice energy and enthalpy of sublimation as descriptors. Salahinejad *et al*. found that these descriptors offered little to the model. This is an unexpected finding, as the general assertion is that lattice interactions play an important role in solubilizing a compound. Here we apply a full range of physically motivated sublimation descriptors, including those used by Salahinejad *et al.* (lattice energy and enthalpy of sublimation), to predict the sublimation terms rather than solubility. If one can predict sublimation free energy to a reasonable accuracy, then in principle it can be combined with any other method for the prediction of hydration free energy and provide a more physically interpretable prediction of solubility than that achievable by a QSPR model for solubility prediction.

We can see from the results presented above that the small SUB-48 dataset contains a number of poor predictions compared to the SUB-158 predictions. We would have preferred to work with the SUB-158 dataset for all aspects of this work, but were unable to due to the absence

of suitable crystallographic data for some molecules. A standard dataset with all thermodynamic terms and polymorphic information available would be a highly prized asset to aid modeling in this area.

The most difficult term to predict, the TΔS of sublimation term, is however consistent between the two data sets. For the SUB-48 dataset, using the PT descriptors, very low if any correlation was found between the predictions and the experimental data. A better correlation is found when CDK descriptors are used, although the correlations are still very low. This is rectified in the SUB-158 prediction by increasing the dataset size. This may suggest the entropy is more sensitive to the dataset than the other thermodynamic parameters.

Given that the free energy is predicted well from CDK descriptors, and appears to be complemented by the inclusion of theoretical chemistry data in SUB-48 predictions, one may expect the constituents of the free energy to be predicted to a similar accuracy. It appears from this analysis that such an assertion is not necessarily the case. Neither the enthalpy nor the entropy of sublimation can here be predicted using CDK descriptors to the same level of accuracy as the free energy of sublimation. In terms of solubility prediction, this suggests that separate QSPR models, one making predictions of the free energy of sublimation and a second one making predictions of the free energy of hydration/solvation, may be a viable way forward to ensure that both steps of the solution process are explicitly accounted for in QSPR predictions of solubility. This would allow limited physical insight to be gained from QSPR predictions of solubility although it is unlikely to be as accurate as a single model predicting solubility.

## Conclusion

We have presented an approximate theoretical method and a number of QSPR methods for the prediction of sublimation thermodynamical properties. We find generally that QSPR

methods can provide a reasonable, although not necessarily quantitatively useful, prediction of each thermodynamic terms over the SUB-158 data set. The TΔS term appears to be the most difficult property to predict over the SUB-158 dataset. For this dataset we find that PLS performs the best over all regression methods, with RF providing the best single prediction of the entropy of sublimation. Additionally, the free energy appears to be the easiest of the three terms to predict. We also note that for both the enthalpy and free energy the importance of the descriptors reduces much more rapidly over the top ten most important compared with those for the entropy.

Over the SUB-48 dataset, we test the application of theoretical chemistry terms as descriptors and their complementarity with CDK 2D descriptors. We find that the enthalpy is marginally better predicted by combining theoretical chemistry methods and QSPR models for the SUB-48 dataset. The free energy of sublimation is reasonably predicted by QSPR models employing 2D descriptors alone, although some improvement is found by combining CDK and PT descriptors. The entropy of sublimation is predicted with poor to modest correlation against experiment and a higher sensitivity to the data provided for training and testing. The entropy in this study appears to be the least predictable of the thermodynamic quantities tested here. We note that whilst the overall predictive accuracy from the SUB-48 dataset is not currently sufficient to improve sublimation prediction, this method may provide a more chemically informed perspective due to the inclusion of theoretical chemistry terms. We hope this may provide a path to improved sublimation thermodynamics predictions. It may be of interest to explore the removal of difficult to predict groups and to produce independent models for different chemical classes. More advanced feature reduction algorithms may also benefit these prediction schemes.

We find that the inclusion of theoretically calculated energetic values as descriptors for the crystalline transition marginally improves the predictive accuracy of the enthalpy and free

energy of sublimation by QSPR models. We therefore recommend the use of similar descriptors, where possible, in future studies. A larger dataset is required to thoroughly test to possible improvements. However, there is clearly much work to be done in improving affordable theoretical methods. The lack of intra-molecular flexibility is a key issue in these models. We have seen that simple corrections for the number of rotatable bonds are insufficient and do not provide any significant improvement in this case, (**Figure S12**) although a generally increasing trend is found between the experimental thermodynamic parameters and the number of rotatable bonds (**Figures S41 – S43)**. For future work, it is possible that flexible models and global energy minimization in the crystal[117] and vacuum structure could provide a significant improvement, especially in terms of entropy. This issue may require more advanced treatment of the multipolar electrostatics, such as fast multipoles[118] or even machine learning models to predict the multipoles allowing for fast updates to the multipolar electrostatics and hence, the inclusion of intra-molecular flexibility where such a multipolar series is convergent.[119]

Finally, there is a need for a standard experimental dataset containing all relevant thermodynamic terms (enthalpy, entropy and free energy), experimental conditions and where possible the polymorph or pseudo-polymorph involved. These data could help greatly in the advancement of this field. We include our own dataset in the Supporting Information, which has been curated from a wide variety of literature[52–105] for use in sublimation thermodynamics prediction.

## ASSOCIATED CONTENT

Full experimental data with references to the original work: Plots of, Enthalpy- Entropy compensation, molecular weight against enthalpy and the effect of including a rotatable bond correction on ΔG of sublimation prediction using DMACRYS: A table of experimental data

and DMACRYS predictions. This material is available free of charge via the Internet at http://pubs.acs.org.

## AUTHOR INFORMATION


### Corresponding Author

* jbom@st-andrews.ac.uk


### Author Contributions

The concepts and ideas in this study were discussed and developed by all authors. Calculations were carried out by JMcD utilizing scripts created by DSP and JMcD. Analysis and discussions of the results were carried out by all authors.


### Funding Sources

JMcD and JBOM would like to thank SULSA for funding. DSP thanks the University of Strathclyde for support through its Strategic Appointment and Investment Scheme.

## ACKNOWLEDGMENT

JMcD, JBOM and TvM thank the EaStCHEM Research Computing Facility and Dr Herbert Früchtl for its upkeep and useful discussions. The authors also gratefully acknowledge a NSCCS computer time grant (CHEM647) and the superb training and advisory service offered through the NSCCS by Dr Alexandra Simperler. JMcD would also like to thank Prof. Graeme Day for additional scripts to calculate the crystalline entropy using DMACRYS and Dr Neetika Nath and Dr Luna De Ferrari for help with the machine learning procedure.


## ABBREVIATIONS

$\Delta G_{sol}$, Free energy of solution; $\Delta G_{sub}$, Free energy of sublimation; $\Delta G_{hyd}$, Free energy of hydration; $\Delta G_{mix}$, Free energy of mixing; $\Delta H_{sub}$,Enthalpy of sublimation; $\Delta S_{sub}$,Entropy of sublimation; $T\Delta S_{sub}$, Temperature multiplied by the entropy of sublimation; 2D, two

dimensional; BCS, Biopharmaceutics Classification System; CDK, Chemistry Development Kit ; CPU, Central Processing Unit ; CSD, Cambridge Structural Database ; PLS, Partial Least Squares (Projection to Latent Structure) ; QSPR, Quantitative Structure–Property Relationship ; RF , Random Forest ; RMSE, Root Mean Square Error ; SMILES , Simplified Molecular Input Line Entry System ; SVM, Support Vector Machine ; TPSA, Topological Surface Area.

KEYWORDS Computational Chemistry, Machine learning, Pharmaceuticals, QSPR, Sublimation, Solubility, Free energy, entropy, enthalpy, crystal structure, polymorph, lattice energy, Random Forest, Support Vector Machines, Partial Least Squares, Thermodynamics.

REFERENCES

(1) Docherty, R.; Pencheva, K.; Abramov, Y. A. Low Solubility in Drug Development: De-Convoluting the Relative Importance of Solvation and Crystal Packing. *J. Pharm. Pharmacol.* **2015**, *67* (6), 847–856.

(2) Perlovich, G. L.; Raevsky, O. A. Sublimation of Molecular Crystals: Prediction of Sublimation Functions on the Basis of HYBOT Physicochemical Descriptors and Structural Clusterization. *Cryst. Growth Des.* **2010**, *10* (6), 2707–2712.

(3) de Klerk, N. J. J.; van den Ende, J. A.; Bylsma, R.; Grančič, P.; de Wijs, G. A.; Cuppen, H. M.; Meekes, H. Q -GRID: A New Method To Calculate Lattice and Interaction Energies for Molecular Crystals from Electron Densities. *Cryst. Growth Des.* **2016**, *16* (2), 662–671.

(4) Perlovich, G. L.; Kurkov, S. V.; Kinchin, A. N.; Bauer-Brandl, A. Thermodynamics of Solutions III: Comparison of the Solvation of (+)-Naproxen with Other NSAIDs. *Eur. J. Pharm. Biopharm.* **2004**, *57* (2), 411–420.

(5) Keiser, D.; Kana'an, A. S. Enthalpy and Entropy of Sublimation of Tetraphenyltin and Hexaphenylditin. Bond Dissociation Energy of Sn-C and Sn-Sn. *J. Phys. Chem.* **1969**, *73* (12), 4264–4269.

(6) Surov, A. O.; Bui, C. T.; Volkova, T. V; Proshin, A. N.; Perlovich, G. L. The Impact of Structural Modification of 1,2,4-Thiadiazole Derivatives on Thermodynamics of Solubility and Hydration Processes. *Phys. Chem. Chem. Phys.* **2015**, *17* (32), 20889–20896.

(7) Salahinejad, M.; Le, T. C.; Winkler, D. A. Aqueous Solubility Prediction: Do Crystal Lattice Interactions Help? *Mol. Pharmaceutics* **2013**, *10* (7), 2757–2766.

(8) Salahinejad, M.; Le, T. C.; Winkler, D. A. Capturing the Crystal: Prediction of Enthalpy of Sublimation, Crystal Lattice Energy, and Melting Points of Organic Compounds. *J. Chem. Inf. Model.* **2013**, *53* (1), 223–229.

(9) Abramov, Y. A. Major Source of Error in QSPR Prediction of Intrinsic Thermodynamic Solubility of Drugs: Solid vs Nonsolid State Contributions? *Mol. Pharmaceutics* **2015**, *12* (6), 2126–2141.

(10) McDonagh, J. L.; Nath, N.; De Ferrari, L.; Van Mourik, T.; Mitchell, J. B. O. Uniting Cheminformatics and Chemical Theory to Predict the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J. Chem. Inf. Model.* **2014**, *54* (3), 844–856.

(11) Palmer, D. S.; McDonagh, J. L.; Mitchell, J. B. O.; Van Mourik, T.; Fedorov, M. V. First-Principles Calculation of the Intrinsic Aqueous Solubility of Crystalline Druglike Molecules. *J. Chem. Theory Comput.* **2012**, *8* (9), 3322–3337.

(12) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41* (5), 1208–1217.

(13) Sirius-analytical. Solubility Definitions http://www.sirius-analytical.com/science/solubility/solubility-definitions (accessed Jun 5, 2015).

(14) McDonagh, J. L. Computing the Aqueous Solubility of Organic Drug-like Molecules and Understanding Hydrophobicity, 2015, http://hdl.handle.net/10023/6534.

(15) Savjani, K. T.; Gajjar, A. K.; Savjani, J. K. Drug Solubility: Importance and Enhancement Techniques. *ISRN Pharm.* **2012**, *2012*, 1–10.

(16) Di, L.; Kerns, E. H.; Carter, G. T. Drug-like Property Concepts in Pharmaceutical Design. *Curr. Pharm. Des.* **2009**, *15* (19), 2184–2194.

(17) Alderson, R. G.; De Ferrari, L.; Mavridis, L.; McDonagh, J. L.; Mitchell, J. B. O.; Nath, N. Enzyme Informatics. *Curr. Top. Med. Chem.* **2012**, *12* (17), 1911–1923.

(18) Sliwoski, G.; Kothiwale, S.; Meiler, J.; Lowe, E. W. Computational Methods in Drug Discovery. *Pharmacol. Rev.* **2014**, *66* (1), 334–395.

(19) Skyner, R. E.; McDonagh, J. L.; Groom, C. R.; van Mourik, T.; Mitchell, J. B. O. A Review of Methods for the Calculation of Solution Free Energies and the Modelling of Systems in Solution. *Phys. Chem. Chem. Phys.* **2015**, *17* (9), 6174–6191.

(20) Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Why Are Some Properties More Difficult to Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P. *J. Chem. Inf. Model.* **2008**, *48* (1), 220–232.

(21) Lusci, A.; Pollastri, G.; Baldi, P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-like Molecules. *J. Chem. Inf. Model.* **2013**, *53* (7), 1563–1575.

(22) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models to Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2006**, *47* (1), 150–158.

(23) Raevsky, O. A.; Polianczyk, D. E.; Grigorev, V. Y.; Raevskaja, O. E.; Dearden, J. C. In Silico Prediction of Aqueous Solubility: A Comparative Study of Local and Global Predictive Models. *Mol. Inform.* **2015**, *34* (6-7), 417–430.

(24) Palmer, D. S.; Mitchell, J. B. O. Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules ? Is Experimental Data Quality the Limiting Factor in Predicting the Aqueous Solubility of Druglike Molecules ? *Mol. Pharmaceutics* **2014**, *11* (8), 2962–2972.

(25) Avdeef, A. Suggested Improvements for Measurement of Equilibrium Solubility-pH of Ionizable Drugs. *ADMET DMPK* **2015**, *3* (2), 84–109.

(26) Palmer, D. S.; Llinas, A.; Inaki, M.; Day, G. M.; Goodman, J. M.; Glen, R. C.; Mitchell, J. B. O. Predicting Intristic Aqueous Solubility by a Thermodynamic Cycle. *Mol. Pharmaceutics* **2008**, *5* (2), 266–279.

(27) Kew, W.; Mitchell, J. B. O. Greedy and Linear Ensembles of Machine Learning Methods Outperform Single Approaches for QSPR Regression Problems. *Mol. Inform.* **2015**, *34* (9), 634–647.

(28) Nigsch, F.; Bender, A.; Van Buuren, B.; Tissen, J.; Nigsch, E.; Mitchell, J. B. O. Melting Point Prediction Employing K-Nearest Neighbor Algorithms and Genetic Parameter Optimization. *J. Chem. Inf. Model.* **2006**, *46* (6), 2412–2422.

(29) Bergström, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. Molecular Descriptors Influencing Melting Point and Their Role in Classification of Solid Drugs. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (4), 1177–1185.

(30) McDonagh, J. L.; van Mourik, T.; Mitchell, J. B. O. Predicting Melting Points of Organic Molecules: Applications to Aqueous Solubility Prediction Using the General Solubility Equation. *Mol. Inform.* **2015**, *34*, 715–724.

(31) Delaney, J. S. Predicting Aqueous Solubility from Structure. *Drug Discov. Today* **2005**, *10* (4), 289–295.

(32) Sumi, T.; Mitsutake, A.; Maruyama, Y. A Solvation-Free-Energy Functional: A Reference-Modified Density Functional Formulation. *J. Comput. Chem.* **2015**, *36* (18), 1359–1369.

(33) Palmer, D. S.; Frolov, A. I.; Ratkova, E. L.; Fedorov, M. V. Towards a Universal Method for Calculating Hydration Free Energies: A 3D Reference Interaction Site Model with Partial Molar Volume Correction. *J. Phys. Condens. Matter* **2010**, *22* (49), 1-9.

(34) Ratkova, E. L.; Fedorov, M. V. Combination of RISM and Cheminformatics for Efficient Predictions of Hydration Free Energy of Polyfragment Molecules: Application to a Set of Organic Pollutants. *J. Chem. Theory Comput.* **2011**, *7* (5), 1450–1457.

(35) Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M. Modelling Organic Crystal Structures Using Distributed Multipole and Polarizability-Based Model Intermolecular Potentials. *Phys. Chem. Chem. Phys.* **2010**, *12* (30), 8478–8490.

(36) Stone, A. J. Distributed Multipole Analysis of Gaussian Wavefunctions GDMA Version 2.2.02, 2005.

(37) Stone, A. J. Distributed Multipole Analysis : Stability for Large Basis Sets Distributed Multipole Analysis : Stability for Large Basis Sets. *Analysis* **2005**, *1* (6), 1128–1132.

(38) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A. *et al.* Gaussian 09, 2009, Revision C.1; Gaussian, Inc.: Wallingford, CT.

(39) Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr. Sect. B Struct. Sci.* **2002**, *58*, 380–388.

(40) Pudipeddi, M.; Serajuddin, A. T. M. Trends in Solubility of Polymorphs. *J. Pharm. Sci.* **2005**, *94* (5), 929–939.

(41) Becke, A. D. Density-Functional thermochemistry.III. The Role of Exact Exchange. *J. Chem. Phys.* **1993**, *98* (7), 5648.

(42) Williams, D. E. Nonbonbonded Potential Parameters Derived from Crystalline Aromatic Hydrocarbons. *J. Chem. Phys.* **1966**, *45* (10), 3770–3778.

(43) Maple, J. R.; Dinur, U.; Hagler, A. T. Derivation of Force Fields for Molecular Mechanics and Dynamics from Ab Initio Energy Surfaces. *Proc. Natl. Acad. Sci* **1988**, *85* (15), 5350–5354.

(44) Williams, D. E.; Hsu, L.-Y. Transferability of Nonbonded Cl...Cl Potential Energy Function to Crystalline Chlorine. *Acta Crystallogr. Sect. A Found. Crystallogr.* **1985**, *41* (3), 296–301.

(45) Day, G.; Price, S.; Leslie, M. Atomistic Calculations of Phonon Frequencies and Thermodynamic Quantities for Crystals of Rigid Organic Molecules. *J. Phys. Chem. B* **2003**, *107* (39), 10919–10933.

(46) Wang, L. The Gas-Phase Thermochemistry of SeFn, SeFn +, and SeFn - (N = 1-6) from Gaussian-3 Calculations. *Int. J. Mass Spectrom.* **2007**, *264* (1), 84–91.

(47) Gavezzotti, A. *Theoretical Aspects and Computer Modeling of the Molecular Solid State*; Gavezzotti, A., Ed.; Wiley and Sons: Chichester, 1998; Vol. 447. 61 -99

(48) Anderson, E.; Veith, G.; Weininger, D. *SMILES: A Line Notation And Computerized Interpreter for Chemical Structures*; US Environmental Protection Agency, Environmental Research Laboratory, 1987.

(49) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 493–500.

(50) Fitzgerald, N. CDK Small Molecule Descriptors http://sysbiolab.bio.ed.ac.uk/wiki/index.php/CDK_Small_Molecule_Descriptors (accessed Mar 20, 2014).

(51) Kuhn, S.; Truszkowski, A.; Masak, C.; Dmitry, K.; Willighagen, E.; Steinbeck, C.; Torrance, G.; Gilman, B.; Berg, A.; Kerssemakers, J. *et al.* cdk 1.5.13 API http://cdk.github.io/cdk/1.5/docs/api/overview-summary.html (accessed Sep 27, 2016).

(52) Perlovich, G. L.; Kurkov, S. V.; Hansen, L. K. R.; Bauer-Brandl, A. Thermodynamics of Sublimation, Crystal Lattice Energies, and Crystal Structures of Racemates and Enantiomers: (+)- and (+/-)-Ibuprofen. *J. Pharm. Sci.* **2004**, *93* (3), 654–666.

(53) Lima Carlos, F. R. A. C.; Rocha, M. A. A.; Schröder, B.; Gomes, L. R.; Low, J. N.; Santos, L. M. N. B. F. Phenylnaphthalenes: Sublimation Equilibrium, Conjugation, and Aromatic Interactions. *J. Phys. Chem. B* **2012**, *116* (11), 3557–3570.

(54) Ribeiro da Silva, M. A. V; Santos, L. M. N. B. F.; Lima, L. M. S. S.; da Silva, M. A. V. Thermodynamic Study of 1,2,3-Triphenylbenzene and 1,3,5-Triphenylbenzene. *J. Chem. Thermodyn.* **2010**, *42* (1), 134–139.

(55) Monte, M. J. S.; Hillesheim, D. M. Thermodynamic Study on the Sublimation of 2-Phenylacetic 4-Phenylbutyric, and 5-Phenylvaleric Acid. *J. Chem. Eng. Data* **2001**, *46* (6), 1601–1604.

(56) Cundall, R. B.; Frank Palmer, T.; Wood, C. E. C. Vapour Pressure Measurements on Some Organic High Explosives. *J. Chem. Soc. Faraday Trans. 1* **1978**, *74*, 1339.

(57) Ribeiro da Silva, M. A. V; Monte, M. J. S.; Santos, L. M. N. B. F. The Design, Construction, and Testing of a New Knudsen Effusion Apparatus. *J. Chem. Thermodyn.* **2006**, *38* (6), 778–787.

(58) Perlovich, G. L.; Volkova, T. V.; Proshin, A. N.; Sergeev, D. Y.; Bui, C. T.; Petrova, L. N.; Bachurin, S. O. Synthesis, Pharmacology, Crystal Properties, and Quantitative Solvation Studies from a Drug Transport Perspective for Three New 1,2,4-Thiadiazoles. *J. Pharm. Sci.* **2010**, *99* (9), 3754–3768.

(59) Palmer, D. S.; Llinàs, A.; Morao, I.; Day, G. M.; Goodman, J. M.; Glen, R. C.; Mitchell, J. B. O.

Predicting Intrinsic Aqueous Solubility by a Thermodynamic Cycle. *Mol. Pharmaceutics* **2008**, *5* (2), 266–279.

(60) Ribeiro da Silva, M. A. V; Amaral, L. M. P. F.; Santos, A. F. L. O. M.; Gomes, J. R. B. Thermochemistry of Nitronaphthalenes and Nitroanthracenes. *J. Chem. Thermodyn.* **2006**, *38* (6), 748–755.

(61) Rocha, M. A. A.; Lima, C. F. R. A. C.; Santos, L. M. N. B. F. Phase Transition Thermodynamics of Phenyl and Biphenyl Naphthalenes. *J. Chem. Thermodyn.* **2008**, *40* (9), 1458–1463.

(62) Ribeiro da Silva, M. A. V; Monte Manuel, J. S.; Ribeiro José, R. Thermodynamic Study on the Sublimation of Succinic Acid and of Methyl- and Dimethyl-Substituted Succinic and Glutaric Acids. *J. Chem. Thermodyn.* **2001**, *33* (1), 23–31.

(63) Vecchio, S.; Brunetti, B. Standard Sublimation Enthalpies of Some Dichlorophenoxy Acids and Their Methyl Esters. *J. Chem. Eng. Data* **2005**, *50* (2), 666–672.

(64) Vecchio, S.; Brunetti, B. Vapor Pressures and Standard Molar Enthalpies, Entropies, and Gibbs Free Energies of Sublimation of 2,4- and 3,4-Dinitrobenzoic Acids. *J. Chem. Thermodyn.* **2009**, *41* (7), 880–887.

(65) Ribeiro da Silva, M. A. V.; Ribeiro da Silva, M. D. M. C.; Lobo Ferreira, A. I. M. C.; Santos, A. F. L. O. M.; Galvão, T. L. P. Experimental Thermochemical Study of 2,5- and 2,6-Dichloro-4-Nitroanilines. *J. Chem. Thermodyn.* **2009**, *41* (10), 1074–1080.

(66) Perlovich, G. L.; Volkova, T. V; Manin, A. N.; Bauer-Brandl, A. Influence of Position and Size of Substituents on the Mechanism of Partitioning: A Thermodynamic Study on Acetaminophens, Hydroxybenzoic Acids, and Parabens. *AAPS PharmSciTech* **2008**, *9* (1), 205–216.

(67) Ribeiro da Silva, M. A. V; Fonseca, J. M. S.; Carvalho, R. P. B. M.; Monte, M. J. S. Thermodynamic Study of the Sublimation of Six Halobenzoic Acids. *J. Chem. Thermodyn.* **2005**, *37* (3), 271–279.

(68) Ribeiro Da Silva, M. A. V; Amaral, L. M. P. F.; Santos, A. F. L. O. M.; Gomes, J. R. B. Thermochemistry of Some Alkylsubstituted Anthracenes. *J. Chem. Thermodyn.* **2006**, *38* (4), 367–375.

(69) Monte, M. J. S.; Hillesheim, D. M. Thermodynamic Study on the Sublimation of the Three Iodobenzoic Acids and of 2-Fluoro- and 3-Fluorobenzoic Acids. *J. Chem. Thermodyn.* **2000**, *32* (12), 1727–1735.

(70) Perlovich, G. L.; Volkova, T. V.; Maimin, A. N.; Bauer-Brandl, A. Extent and Mechanism of Solvation and Partitioning of Isomers of Substituted Benzoic Acids: A Thermodynamic Study in the Solid State and in Solution. *J. Pharm. Sci.* **2008**, *97* (9), 3883–3896.

(71) Monte, M. J. S.; Hillesheim, D. M. Thermodynamic Study on the Sublimation of Six Methylnitrobenzoic Acids. *J. Chem. Thermodyn.* **2001**, *33* (1), 103–112.

(72) Ribeiro da Silva, M. A. V; Santos, A. F. L. O. M. Standard Molar Enthalpies of Formation and of Sublimation of 2-Thiophenecarboxamide and 2-Thiopheneacetamide. *J. Chem. Thermodyn.* **2008**, *40* (2), 166–173.

(73) Monte, M. J. S.; Hillesheim, D. M. Thermodynamic Study on the Sublimation of 3-Phenylpropionic Acid and of Three Methoxy-Substituted 3-Phenylpropionic Acids. *J. Chem. Thermodyn.* **2001**, *33* (8), 837–847.

(74) Perlovich, G. L.; Volkova, T. V.; Bauer-Brandl, A. Towards an Understanding of the Molecular Mechanism of Solvation of Drug Molecules: A Thermodynamic Approach by Crystal Lattice Energy, Sublimation, and Solubility Exemplified by Hydroxybenzoic Acids. *J. Pharm. Sci.* **2006**,

*95* (7), 1448–1458.

(75) Monte, M. J. S.; Santos, L. M. N. B. F.; Fonseca, J. M. S.; Sousa, C. A. D. Vapour Pressures, Enthalpies and Entropies of Sublimation of Para Substituted Benzoic Acids. *J. Therm. Anal. Calorim.* **2010**, *100* (2), 465–474.

(76) de Kruif, C. G.; Voogd, J.; Offringa, J. C. A. Enthalpies of Sublimation and Vapour Pressures of 14 Amino Acids and Peptides. *J. Chem. Thermodyn.* **1979**, *11* (7), 651–656.

(77) Perlovich, G. L.; Strakhova, N. N.; Kazachenko, V. P.; Volkova, T. V.; Tkachev, V. V.; Schaper, K. J.; Raevsky, O. a. Sulfonamides as a Subject to Study Molecular Interactions in Crystals and Solutions: Sublimation, Solubility, Solvation, Distribution and Crystal Structure. *Int. J. Pharm.* **2008**, *349* (1-2), 300–313.

(78) Perlovich, G. L.; Ryzhakov, A. M.; Tkachev, V. V; Hansen, L. K. Sulfonamide Molecular Crystals: Thermodynamic and Structural Aspects. *Cryst. Growth Des.* **2011**, *11* (4), 1067–1081.

(79) Perlovich, G. L.; Tkachev, V. V.; Strakhova, N. N.; Kazachenko, V. P.; Volkova, T. V.; Surov, O. V.; Schaper, K. J.; Raevsky, O. a. Thermodynamic and Structural Aspects of Sulfonamide Crystals and Solutions. *J. Pharm. Sci.* **2009**, *98* (12), 4738–4755.

(80) Vecchio, S.; Tomassetti, M. Vapor Pressures and Standard Molar Enthalpies, Entropies and Gibbs Energies of Sublimation of Three 4-Substituted Acetanilide Derivatives. *Fluid Phase Equilib.* **2009**, *279* (1), 64–72.

(81) Monte, M. J. S.; Almeida, A. R. R. P.; Ribeiro da Silva, M. A. V. Thermodynamic Study of the Sublimation of Eight 4-N-Alkylbenzoic Acids. *J. Chem. Thermodyn.* **2004**, *36* (5), 385–392.

(82) Ribeiro da Silva, M. A. V; Lima, L. M. S. S.; Moreno, A. R. G.; Ferreira, A. I. M. C. L.; Gomes, J. R. B. Combined Experimental and Computational Thermochemistry of Isomers of Chloronitroanilines. *J. Chem. Thermodyn.* **2008**, *40* (2), 155–165.

(83) Cox, J. D.; Gundry, H. A.; Harrop, D.; Head, A. J. Thermodynamic Properties of Fluorine Compounds 9. Enthalpies of Formation of Some Compounds Containing the Pentafluorophenyl Group. *J. Chem. Thermodyn.* **1969**, *1* (1), 77–87.

(84) Perlovich, G. L.; Hansen, L. K.; Volkova, T. V.; Mirza, S.; Manin, A. N.; Bauer-Brandl, A. Thermodynamic and Structural Aspects of Hydrated and Unhydrated Phases of 4-Hydroxybenzamide. *Cryst. Growth Des.* **2007**, *7* (12), 2643–2648.

(85) Colomina, M.; Jiménez, P.; Roux, M. V.; Turrión, C. Thermochemical Properties of Benzoic Acid Derivatives VII. Enthalpies of Combustion and Formation of the O-, M-, and P-Methoxybenzoic Acids. *J. Chem. Thermodyn.* **1978**, *10* (7), 661–665.

(86) Monte, M. J. S.; Sousa, C. A. D. Vapor Pressures and Phase Changes Enthalpy and Gibbs Energy of Three Crystalline Monomethyl Benzenedicarboxylates. *J. Chem. Eng. Data* **2005**, *50* (6), 2101–2105.

(87) Ribeiro Da Silva, M. A. V; Matos, M. A. R.; Monte, M. J. S.; Hillesheim, D. M.; Marques, M. C. P. O.; Vieira, N. F. T. G. Enthalpies of Combustion, Vapour Pressures, and Enthalpies of Sublimation of Three Methoxy-Nitrobenzoic Acids. Vapour Pressures and Enthalpies of Sublimation of the Three Nitrobenzoic Acids. *J. Chem. Thermodyn.* **1999**, *31* (11), 1429–1441.

(88) Perlovich, G. L.; Volkova, T. V; Bauer-Brandl, A. Towards an Understanding of the Molecular Mechanism of Solvation of Drug Molecules: A Thermodynamic Approach by Crystal Lattice Energy, Sublimation, and Solubility Exemplified by Paracetamol, Acetanilide, and Phenacetin. *J. Pharm. Sci.* **2006**, *95* (10), 2158–2169.

(89) Vecchio, S.; Brunetti, B. Vapor Pressures and Standard Molar Sublimation Enthalpies of Three

6-Methylthio-2,4-Di(alkylamino)-1,3,5-Triazine Derivatives: Simetryn, Ametryn, and Terbutryn. *J. Chem. Eng. Data* **2007**, *52* (5), 1585–1594.

(90) Perlovich, G. L.; Kurkov, S. V; Kinchin, A. N.; Bauer-Brandl, A. Solvation and Hydration Characteristics of Ibuprofen and Acetylsalicylic Acid. *AAPS J.* **2004**, *6* (1), 22–30.

(91) Perlovich, G. L.; Volkova, T. V.; Bauer-Brandl, A. Thermodynamic Study of Sublimation, Solubility, Solvation, and Distribution Processes of Atenolol and Pindolol. *Mol. Pharmaceutics* **2007**, *4* (6), 929–935.

(92) Aihara, A. Estimation of the Energy of Hydrogen Bonds Formed in Crystals. I. Sublimation Pressures of Some Organic Molecular Crystals and the Additivity of Lattice Energy. *Bull. Chem. Soc. Jpn.* **1959**, *32* (11), 1242–1248.

(93) Perlovich, G. L.; Kurkov, S. V; Kinchin, A. N.; Bauer-Brandl, A. Thermodynamics of Solutions IV: Solvation of Ketoprofen in Comparison with Other NSAIDs. *J. Pharm. Sci.* **2003**, *92* (12), 2502–2511.

(94) Ribeiro da Silva, M. A. V; Monte, M. J. S. The Construction, Testing and Use of a New Knudsen Effusion Apparatus. *Thermochim. Acta* **1990**, *171*, 169–183.

(95) Nitta, Isamu and Seki, S. Vapor Pressures of Molecular Crystals: Aromatic Nitro Compoundstle. *J Chem Soc Jpn, Pure Chem Sec (Nippon Kagaku Zasshi)* **1950**, *71*, 378.

(96) Perlovich, G. L.; Surov, A. O.; Hansen, L. K.; Bauer-Brandl, A. Energetic Aspects of Diclofenac Acid in Crystal Modifications and in Solutions - Mechanism of Solvation, Partitioning and Distribution. *J. Pharm. Sci.* **2007**, *96* (5), 1031–1042.

(97) Kurkov, S. V.; Perlovich, G. L. Thermodynamic Studies of Fenbufen, Diflunisal, and Flurbiprofen: Sublimation, Solution and Solvation of Biphenyl Substituted Drugs. *Int. J. Pharm.* **2008**, *357* (1-2), 100–107.

(98) Victoria Roux, M.; Jimenez, P.; Dávalos, J. Z.; Turrión, C.; Afeefy, H. Y.; Liebman, J. F. Enthalpies of Formation of Methyl Benzenecarboxylates. *J. Chem. Soc. Faraday Trans.* **1998**, *94* (7), 887–890.

(99) Surov, A. O.; Terekhova, I. V.; Bauer-Brandl, A.; Perlovich, G. L. Thermodynamic and Structural Aspects of Some Fenamate Molecular Crystals. *Cryst. Growth Des.* **2009**, *9* (2), 3265–3272.

(100) Perlovich, G. L.; Surov, A. O.; Bauer-Brandl, A. Thermodynamic Properties of Flufenamic and Niflumic Acids-Specific and Non-Specific Interactions in Solution and in Crystal Lattices, Mechanism of Solvation, Partitioning and Distribution. *J. Pharm. Biomed. Anal.* **2007**, *45* (4), 679–687.

(101) Vecchio, S. Vapor Pressures and Standard Molar Enthalpies, Entropies and Gibbs Energies of Sublimation of Two Hexachloro Herbicides Using a TG Unit. *Thermochim. Acta* **2010**, *499* (1-2), 27–33.

(102) Ribeiro da Silva, C, M. D. D. M.; Ribeiro da Silva, M. A. V; Freitas, V. L. S.; Roux, M. V.; Jiménez, P.; Temprado, M.; Dávalos, J. Z.; Cabildo, P.; Claramunt, R. M.; Elguero, J. Structural Studies of Cyclic Ureas: 1. Enthalpies of Formation of Imidazolidin-2-One and N,N'-trimethyleneurea. *J. Chem. Thermodyn.* **2008**, *40* (3), 386–393.

(103) Ribeiro da Silva, M. A. V; Santos, L. M. N. B. F.; Lima, L. M. S. S. Standard Molar Enthalpies of Formation and of Sublimation of the Terphenyl Isomers. *J. Chem. Thermodyn.* **2008**, *40* (3), 375–385.

(104) Perlovich, G. L.; Strakhova, N. N.; Kazachenko, V. P.; Volkova, T. V.; Tkachev, V. V.; Schaper, K. J.; Raevsky, O. a. Studying Thermodynamic Aspects of Sublimation, Solubility and Solvation

Processes and Crystal Structure Analysis of Some Sulfonamides. *Int. J. Pharm.* **2007**, *334* (1-2), 115–124.

(105) Monte, M. J. S.; Santos, L. M. N. B. F.; Fulem, M.; Fonseca, J. M. S.; Sousa, C. a D. New Static Apparatus and Vapor Pressure of Reference Materials: Naphthalene, Benzoic Acid, Benzophenone, and Ferrocene. *J. Chem. Eng. Data* **2006**, *51* (2), 757–766.

(106) Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 468–481.

(107) Svetnik, V.; Liaw, A.; Tong, C.; Christopher Culberson, J.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947–1958.

(108) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*, Revised Ed.; Springer: Dordrecht, 2007.

(109) Kuhn, M. Predictive Modeling with R and the caret Package https://www.r-project.org/nosvn/conferences/useR-2013/Tutorials/kuhn/user_caret_2up.pdf (accessed May 5, 2014).

(110) Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; The R Core Team; Benesty, M.; Lescarbeau, R.; Ziem, A.; Scrucca, L.; Tang, Y.; Candan, C. Package caret, **2015**, 75–77.

(111) McDonagh, J. L.; Nath, N.; De Ferrari, L.; Mitchell, J. B. O. ML-Algorithm and SUB-158 dataset http://chemistry.st-andrews.ac.uk/staff/jbom/group/sublimation_scripts_data.zip (accessed Mar 22, 2016).

(112) Nath, N.; De Ferrari, L.; McDonagh, J. L. Machine Learning R scripts - RF - SVM - PLS https://github.com/Jammyzx1/R-ML-RF-SVM-PLS (accessed Sep 27, 2016).

(113) Kier, L. B.; Hall, L. H. An Electrotopological-State Index for Atoms in Molecules. *Pharm. Res.* **1990**, *7* (8), 801–807.

(114) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43* (20), 3714–3717.

(115) Keshavarz, M. H.; Bashavard, B.; Goshadro, A.; Dehghan, Z.; Jafari, M. Prediction of Heats of Sublimation of Energetic Compounds Using Their Molecular Structures. *J. Therm. Anal. Calorim.* **2015**, *120* (3), 1941–1951.

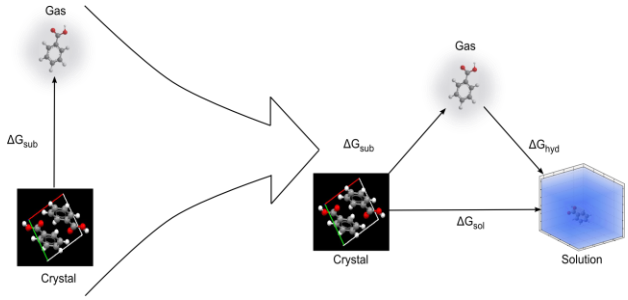
(116) Ouvrard, C.; Mitchell, J. B. O. Can We Predict Lattice Energy from Molecular Structure? *Acta Crystallogr. Sect. B Struct. Sci.* **2003**, *59* (5), 676–685.

(117) Pyzer-Knapp, E. O.; Thompson, H. P. G.; Schiffmann, F.; Jelfs, K. E.; Chong, S. Y.; Little, M. A.; Cooper, A. I.; Day, G. M. Predicted Crystal Energy Landscapes of Porous Organic Cages. *Chem. Sci.* **2014**, *5* (6), 2235.

(118) Mathias, G.; Egwolf, B.; Nonella, M.; Tavan, P. A Fast Multipole Method Combined with a Reaction Field for Long-Range Electrostatics in Molecular Dynamics Simulations: The Effects of Truncation on the Properties of Water. *J. Chem. Phys.* **2003**, *118* (24), 10847.

(119) Popelier, P. L. A. QCTFF: On the Construction of a Novel Protein Force Field. *Int. J. Quantum Chem.* **2015**, *115* (16), 1005–1011.

Insert Table of Contents Graphic and Synopsis Here


Gas
ΔG_sub
Crystal
Gas
ΔG_sub
ΔG_hyd
ΔG_sol
Crystal
Solution