# Prediction of hydrate and solvate formation using statistical models

*Khaled Takieddin, Yaroslav Z. Khimyak, and László Fábián**

School of Pharmacy, University of East Anglia, Norwich, UK

Solvate, Hydrate, Molecular descriptors, Chemoinformatics, Statistics, Logistic regression, Prediction, Predictive models.

## Abstract

Novel, knowledge-based models for the prediction of hydrate and solvate formation are introduced, which require only the molecular formula as input. A dataset of more than 19,000 organic, non-ionic and non-polymeric molecules was extracted from the Cambridge Structural Database. Molecules that formed solvates were compared with those that did not using molecular descriptors and statistical methods, which allowed the identification of chemical properties that contribute to solvate formation. The study was conducted for five types of solvates: ethanol, methanol, dichloromethane, chloroform and water solvates. The identified properties were all related to the size and branching of the molecules and to the hydrogen bonding ability of the molecules. The corresponding molecular descriptors were used to fit logistic regression models to predict the probability of any given molecule to form a solvate. The established models were

able to predict the behavior of ~80% of the data correctly using only two descriptors in the predictive model.

## 1. Introduction

Pharmaceutical processing steps during manufacturing can lead to an unexpected change in the crystal form of materials.[1] One of the common changes is the inclusion of a solvent into the crystal structure of the drug, i.e. solvate formation. It was estimated that 33% of organic compounds have the ability to form hydrates, while about 10% of them are able to form solvates with organic solvents.[2]

Solvate formation has many implications in the pharmaceutical industry, because it affects the physico-chemical properties of materials, such as their density, melting point and dissolution rate, which in turn can influence their manufacturability and pharmacokinetic properties.[3] The unexpected formation of solvates can thus lead to unpredictable behavior of the drug.

From a more optimistic point of view, the different physical and chemical properties of the hydrate and solvate forms can be utilized to alter the rate of drug release or to stabilize the formulation. There are many examples of drugs that are formulated as a hydrate form, such as cephalexin, cefaclor, ampicillin and theophylline.[4-5] Hydrates (water solvates) are of special concern, because they occur more frequently than other solvates. Another factor that makes hydrates particularly important is the fact that water is a non-toxic solvent. In 2010 the number of hydrate structures from organic, organometallic and coordination compounds in the Cambridge Structural Database was 49,283 out of the total of 443,505 structures in the CSD, which is about 11% of the entries in the database.[2] One of the few examples of a marketed solvate is the HIV protease inhibitor indinavir, which was formulated as the sulfate salt ethanol solvate in order to improve the stability and the bioavailability of the drug.[6]

Although factors affecting solvate and hydrate formation have been investigated previously and predictions were made for specific drugs,[7,8] the general prediction of solvate formation – similar to the prediction of other solid forms – is still largely an unresolved problem.[9] Currently, in order to avoid unexpected structural transformations (such as hydrate and solvate forms) in the pharmaceutical industry, high-throughput crystallization experiments are conducted to obtain all possible solid forms of a drug.[10] Although this method is convenient for screening possible solid forms, it still has some disadvantages. For example, it is never certain that all possible phase transitions have been identified. It is also very difficult to explain why these phase transitions happen. Another disadvantage is the necessity of having the actual material.

The current ability to predict crystal structures can be illustrated by the crystal structure prediction (CSP) blind tests organized at the Cambridge Crystallographic Data Centre.[11] The latest blind test, which was conducted for six molecules, showed that the crystal structures of molecules of different properties (small, rigid, flexible etc.) can be predicted reliably, but among them a hydrate was deemed one of the most challenging structures to predict. This type of predictive methodology is associated with high computational cost and requires a high level of expertise in molecular modelling.

Another approach is relying on previously conducted experiments. The library of results from screening experiments can be analyzed in order to find a trend or a pattern among the data, which helps in making general conclusions and may allow prediction of the outcomes of future experiments. When this approach is used, two important aspects arise: firstly, the source of information used for the study and secondly, the choice of suitable methods for the analysis.

In order to reliably identify a weak trend in a set of experiments, large amounts of data are required. Databases which aggregate data about previous experiments can be used as an easy way to access the desired information. The huge growth in the size and availability of databases has facilitated their use in the environmental, medical and social sciences.[12-14] Similarly, chemical databases such as the Cambridge Structural Database (CSD),[15] can be used to analyze structural data and draw chemical conclusions.[16]

A number of investigations regarding hydrate and the solvate formation using the CSD have been conducted.[17-20] These identified correlations between hydrate/solvate formation and the possibility of strong, specific hydrogen bonds with the solvent, as well as the overall hydrogen bonding functionality and polarity of the molecules. They also showed some evidence of solvate formation improving the close packing efficiency of large molecules.[19] However, no attempt was made to use these correlations to predict solvate formation.

Data mining techniques (statistics, artificial intelligence and machine learning) can be applied to develop predictive models from large datasets.[21] Examples of predictions using these methods can be found in different research areas,[22-24] including materials science.[16, 25] For instance, the use of machine learning methods for solvate formation has been demonstrated by Johnston *et al.*,[26], who identified three new carbamazepine solvates using a Random Forest[27] classification of 65 solvents.

In this study, we aim to identify molecular properties that are associated with solvate formation in different solvents and develop predictive models using the data mining techniques mentioned earlier and data from the 2014 edition of the CSD. Although over 300 solvents are represented in the CSD,[28] only five solvents will be discussed in this article. Two alcohols, ethanol and methanol, and two apolar chlorinated solvents, dichloromethane and chloroform were selected.

These solvents were chosen to represent two different classes of solvents, so that the ability of the proposed method to distinguish between solvents and solvent classes can be assessed. Water was included as the most abundant solvent in organic crystals.[29] The five solvents chosen showed a large number of hits in the CSD both among solvates and as recrystallization solvents of non-solvated forms. The abundance of available data is important, as it increases the reliability of the resulting models.

The molecular properties were studied via molecular descriptors. These are numerical attributes that are calculated from chemical structures and represent information about them. The properties that they describe numerically range from conceptually simple (e.g. the van der Waals volume of a molecule) to complex ones (e.g. eigenvalues of matrices representing atom-atom connectivity in the molecule).[30] Determining which descriptors contribute to solvate formation will allow us to predict the probability of solvate formation in crystallization experiments of any molecule. This can provide a guide in choosing the right solvent during the development of formulations and manufacturing processes.

## 2. Methods

**2.1 CSD data extraction.** The Cambridge Structural Database (CSD),[15] which currently contains over 700,000 crystal structures, was used as the source of information for this project. The Conquest software was used to search through this database.[31] Two groups of structures were extracted from the CSD database, solvate-forming and non-solvate forming ones. The search for both groups was restricted to entries that are organic, non-polymeric and non-ionic compounds. Limiting the search to this group of molecules helped to avoid the influence of ionic interactions between the molecules in the study.

Solvate-forming structures were identified as having two different chemical entities in the recorded structure, with one of them being the appropriate solvent. The non-solvate-forming group was defined through the recorded use of the solvent under investigation as the recrystallization solvent along with the presence of only one non-solvent chemical entity in the crystal structure.

Each of the extracted structures was saved into a separate file. These files were then processed by custom-made programs to extract a unique non-solvent molecule from each structure (the corresponding Perl and bash scripts are available from the corresponding author on request). 4885 molecular descriptors were calculated for each molecule using the Dragon software.[32] The types of descriptors calculated by Dragon are given in Table S1, Supporting Information, along with examples and references. These descriptors were subsequently analyzed using the R statistical language.[33]

**2.2 Significance testing.** The aim of this step was to identify which descriptors can classify the data into the solvate forming and non-solvate forming groups. The test used was the Wilcoxon signed-rank test. This non-parametric test[34] was used to compare the solvate and the non-solvate forming groups for each solvent and find which descriptors show a statistically significant difference between them, at a *p*-value of 0.05. The test assumes the null hypothesis to be that the two datasets come from the same population and then finds the probability of this hypothesis, *i.e.* the *p*-value. When the *p*-value is less than 0.05, this means the observations support the assumption that the null hypothesis is wrong with at least 95% probability.[35]

**2.3 Machine learning methods.** The descriptors identified in the previous step were then used to fit predictive models. Different pattern recognition techniques, such as artificial neural networks, support vector machines and logistic regression were tested.[36-38] These methods were used to classify the molecules according to their ability to form solvates depending on the molecular descriptors values. The use of logistic regression gave models with superior predictive ability in all cases. For this reason, it is going to be the method discussed in this paper.

*Logistic regression***.** Logistic regression is a binomial classification system, which can take multiple descriptors into account, each having a different weight. The final result is always a value between 0 and 1, representing the probability of an event to happen. The probability is calculated using Equation (1):

$$x = \frac{1}{1 + e^{-(\beta_0 + \beta_i x_i + \cdots + \beta_n x_n)}} \tag{1}$$

where *x* is the probability of an event to happen, $\beta_0$ is the intercept, $\beta_i$ is the coefficient of the *i*th predictor variable, $x_i$ is the *i*th predictor variable and n is the number of predictor variables in the model.

**2.4 Model evaluation.** *Average weighted MSE:* After the predictive models were fitted using logistic regression, their evaluation took place using the average mean squared error (*MSE*) of the 10-fold cross validation, weighted by the sample size of each fold. The MSE of each fold can be calculated using Equation (2):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \tag{2}$$

where $\hat{y}_i$ is the estimated value from the model, $y_i$ is the real value (0 for solvates or 1 for non-solvates) and $n$ is the number of data points. The *average weighted MSE* can be calculated using Equation (3):

$$Average\ Weighted\ MSE\ = \sum_{i=1}^{k} \frac{N_k}{N} MSE_k \tag{3}$$

where $k$ is the number of folds (10), $N_k$ is the sample size in the $k^{\text{th}}$ fold, $N$ is the total number of molecules, and $MSE_k$ is the *MSE* value of the k$^{\text{th}}$ fold. Due to the large sample sizes used in the analysis, the weighting has minimal effect on the results. The factor $\frac{N_k}{N}$ will have a value very close to 0.1 for each fold, even if the number of molecules is not divisible by 10. For simplicity, the *average weighted MSE* calculated by the software is going to be referred to as *MSE* from this point onwards.

To calculate the *MSE*, the dataset is randomly partitioned into a training set, which is used to fit the model and a test set, which is used to evaluate the model performance. The 10-fold cross validation randomly splits the dataset into 10 parts, where 9 parts are used for model fitting and 1 part is used for testing the model. This process is repeated 10 times, which means that all the points in the dataset were used for both fitting the model and testing it. Little variation between the samples shows that the models being fit are robust. The *MSE* value was used as a method for the selection of the best predictive model. This method has the advantage of incorporating both the variance and the bias of the estimator terms.

*AUC:* Another statistical estimate that was used to compare the models was the area under the Receiver Operating Characteristic (ROC) curve or *AUC*. This area represents the probability that a randomly selected positive instance will be ranked more positive than a randomly selected negative one.[39]

*AIC:* Akaike information criterion was also used to measure the relative quality of the fitted models.[40] This criterion aids in deciding how many descriptors to include in the predictive model. It works by giving a penalty for adding variables to the model. The penalty helps avoiding over-fitting of the model. The *AIC* is calculated using Equation (4):

$$AIC = 2k - 2\ln(L) \tag{4}$$

where $k$ is the number of parameters (variables) in the model and $L$ is the maximized value of the likelihood function of the fitted model.[41]

## 3. Results and discussion

**3.1. Data extraction and significance testing.** Using the search criteria mentioned earlier, 19,010 crystal structures were extracted from the CSD. The extracted dataset consisted of 9162 solvate and 9848 non-solvate structures. The breakdown of the data by crystallization solvent is shown in Table 1. Figures S15 to S18 (Supporting Information) illustrate the distribution of molecular weight, donor and acceptor count and LogP values of molecules in each dataset.

**Table 1.** The number of solvate and non-solvate structures in each solvent

| **Solvent** | **Number of structures** | **Solvates** | **Non-solvates** |
|---|---|---|---|
| Ethanol | 4895 | 689 | 4206 |
| Methanol | 4366 | 1518 | 2848 |
| Dichloromethane | 2761 | 1464 | 1297 |
| Chloroform | 2556 | 1363 | 1193 |
| Water | 4432 | 4128 | 304 |
| Total | 19010 | 9162 | 9848 |

A total of 4885 molecular descriptors (variables) were calculated for each of the solvate and non-solvate molecules using Dragon. This calculation yielded 10 datasets: one set of solvate-forming and one set of non-solvate forming molecules for each of the five solvents. Each dataset of solvate-forming molecules was compared with the corresponding set of non-solvate forming molecules to find the molecular descriptors that correlate with solvate formation in this specific solvent.

The first step in the comparison was the Wilcoxon signed-rank test. Each of the calculated descriptors was tested for having a significant difference between the solvate and the non-solvate forming groups using this test. It was conducted in the R language, at a *p*-value of 0.05. This comparison took place on a descriptor-by-descriptor basis.

Figure 1 shows an example of the significance test for two descriptors: the nAT descriptor, which is the number of atoms in a molecule, and the insignificant O% descriptor, which is the percentage of oxygen atoms among the non-hydrogen atoms in the molecule.

Over 2850 descriptors showed a significant difference between the solvate-forming and the non-solvate-forming groups in each tested solvent. Having over 2850 significant descriptors in each solvent dataset was not really meaningful. In order to select the descriptors that are the best (among those) in showing the difference between the solvate-forming and the non-solvate-forming datasets, further statistical investigation was undertaken.

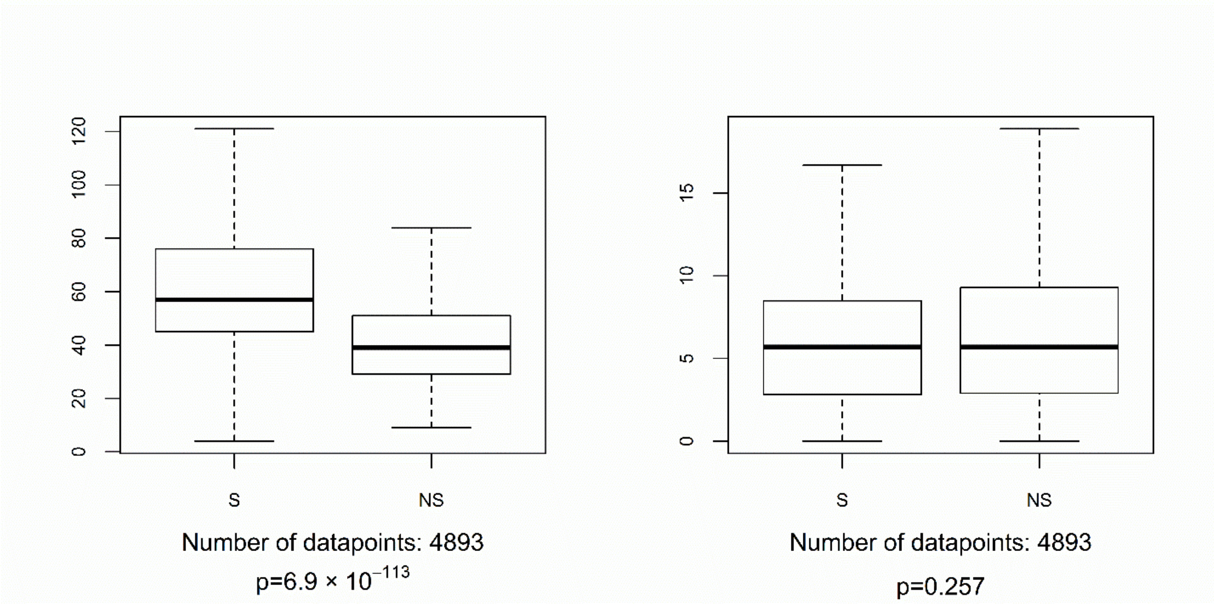


**Figure 1.** Box plot representation of the distribution of two variables in the ethanol dataset. The nAT descriptor (left) shows a significant difference between solvate (S) and non-solvate (NS) molecules, while the O% descriptor (right) does not.

### 3.2 Variable selection

***3.2.1 Single-variable models:*** This approach was based on fitting a logistic regression model of the data using one descriptor at a time. Each model was then validated using a 10-fold cross validation method. The model with the best performance was selected on the basis of the *MSE* value of the 10-fold cross-validation. The *AIC* of the model and the area under the ROC curve for each model were also calculated. The descriptors that were used to fit the logistic regression models were limited to the ones that showed a significant difference between the solvate and the non-solvate groups.

As it has been in shown in Table 1, the number of the solvate and the non-solvate-forming molecules in the extracted data was not even. This imbalance between the two groups can result in predictive models that are biased towards the larger group. Before the models were fitted, samples with equal number of both groups had to be obtained. This means some molecules were removed from the larger dataset. Subsetting the larger dataset to bring the samples to equal sizes took place using random sampling. In order to minimize errors arising from this sampling process, 10 equal size samples were taken based on random seed s. The seeds that were used for random sampling were recorded to ensure the reproducibility of the results.

Each of the 10 samples was tested separately using the method mentioned at the beginning of this section. The descriptor that turned out to have the lowest *MSE* value in most of the 10 samples was considered as the best descriptor for the classification of the data.

The best models fitted using a single molecular descriptor showed a mean standard error (*MSE*) between 0.149 and 0.21 in all five solvents (Table S5, Supporting Information) and therefore had a good predictive ability. They have also shown little variation between the 10 samples used for cross validation. The best single-variable models were related to the so called spectral moment descriptors in each solvent. These descriptors are discussed later in this paper.

***3.2.2 Two-variable models****:* In an attempt to improve the predictive ability of these models, combinations of two descriptors were used to fit logistic regression models. Similarly to the single-descriptor models, only the descriptors that showed significant difference in the Wilcoxon test were considered (more than 2850). The selection of the best model that utilizes two variables required fitting a model with each possible combination of two descriptors. This means that over 4 million models per solvent were fitted. Ten equal size samples were also used for fitting unbiased models. Each of the fitted models was cross-validated using 10-fold cross validation and the *MSE* of each model was recorded. This gives that a total of more than 400 million models were fitted and the best among them were selected. These analyses were programmed in

R[33] and executed using the High Performance Computing Cluster at the University of East Anglia, where the processes were split between 160 dedicated cores. The models for some, but not all solvents improved relative to the single-variable models, as indicated by the reduction of the *MSE* values into the range of 0.145 to 0.184 (Table S6, Supporting Information).

***3.2.3 Three-variable models:*** Since the addition of a second variable improved the *MSE*, *AIC* and *AUC* values of the best-performing models, it was anticipated that the addition of a third descriptor to the models would improve the predictive ability further. The addition of a third descriptor to the models using the same exhaustive approach and the same number of variables (over 2850 per solvent) was not feasible, due to the large number of possible three-variable combinations. Alternatively, the addition of a third variable to the best two-variable model in each solvent was tested, with selection of the best model among the three-variable models based on their *AIC* value. This criterion evaluates the relative quality of models based on the balance between the goodness of fit of the models and their complexity. Therefore, the use of the *AIC* shows whether the addition of a new descriptor to the two-variable models provides significant new information. As carried out previously; the fitting process was repeated 10 times using equally sized subsets of the data. Cross-validation was not deemed essential for this analysis. This is because Stone has shown that the AIC criterion is asymptomatically equivalent to the leave-one-out cross validation.[42] Consequently, a total of about 30,000 models were fitted and the one with the lowest *AIC* value for each solvent was selected. Surprisingly, the amount of information that was added by the third descriptor was very little.

The *MSE* values of the new models fell between 0.142 and 0.184, showing that the addition of the third descriptor to the models did not increase their accuracy or the confidence of the predictions (Table S7, Supporting Information). This means that no significant extra information related to hydrate/solvate formation can be obtained from the calculated descriptors. In order to illustrate the performance of the one, two and three-variable models, a superimposition of the ROC curves of each ethanol model is shown in Figure 2.

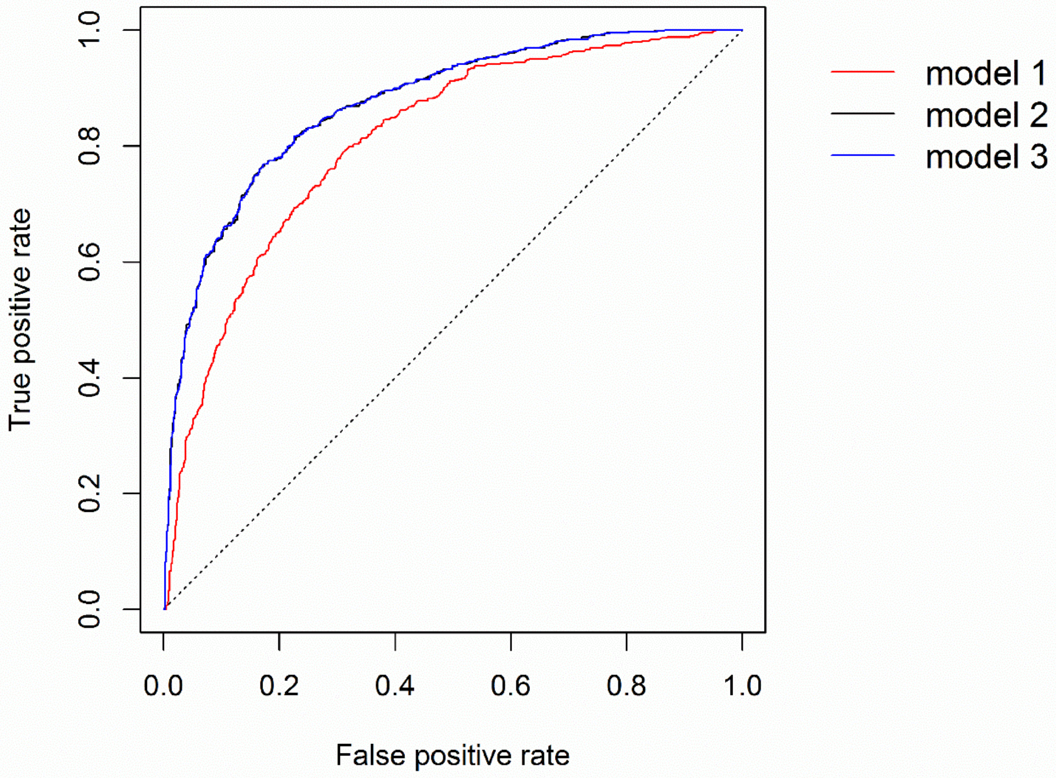


**Figure 2.** ROC curve of the best models utilizing one, two and three variables (model1, model2 and model3, respectively) to predict ethanol solvate formation (sample size: 1377). Model2 and model3 are almost perfectly superimposed due to the small effect of the addition of the third descriptor to model2. Similar representation of the models for the other solvents are given in the Supporting Information (Figures S2-S5)

**3.3 Discussion of the best models.** Although the two-variable models perform significantly better than the one-variable models, the three-variable models do not seem to improve the predictions in any of the data sets. For this reason, only the two-variable models are going to be discussed in detail. A summary of the two-variable models and their performance is shown in Table 2.

**Table 2.** The average performance of the two-variable models over 10 samples in the 5 solvents.

| Model | Descriptors | Intercept | Descriptor 1 coefficient | Descriptor 2 coefficient | No. of data points | *MSE* | *AIC* | *AUC* |
|---|---|---|---|---|---|---|---|---|
| Ethanol | *AVS_H2* + *nHDon* | 15.939 | -3.817 | -0.861 | 1377 | 0.149 | 1283 | 0.868 |
| Methanol | *TRS* + *nHDon* | 2.808 | -0.084 | -0.612 | 3035 | 0.180 | 3268 | 0.810 |
| Dichloromethane | *SM3_H2* + *Hy* | 15.459 | -3.314 | -0.664 | 2592 | 0.146 | 2386 | 0.871 |
| Chloroform | *SM3_H2* + *H-050* | 14.744 | -3.051 | -0.384 | 2384 | 0.148 | 2212 | 0.867 |
| Water | *πID* + *Mor05u* | 4.672 | -0.424 | 0.327 | 607 | 0.159 | 587 | 0.846 |

***3.3.1 The ethanol model:*** The best two-variable model for ethanol utilizes the *AVS_H2* and *nHDon* descriptors. *AVS_H2* is a descriptor derived from the reciprocal squared topological distance matrix, and it is calculated by taking the natural logarithm of the average of the sum of the entries in each row of the matrix.[43] An example of calculating the reciprocal squared topological distance matrix is shown in Supporting Information (Figure S1, Tables S1-S3). The value of the *AVS_H2* descriptor is directly related to molecular size and branching of the molecular graph. The larger the molecule or the more branched it is, the larger this value becomes.

Although this descriptor can be calculated by a computer in a fraction of a second for any given molecule, it would be impractical to calculate it manually. Moreover, the descriptor value is not easily estimated by looking at the molecular structure. In order to give a more intuitive value, models based on closely related descriptors were tested. It turns out that the number of rings in the molecule (*nCIC*) is highly correlated (r = 0.87) to *AVS_H2*. This is a logical correlation as *AVS_H2* incorporates information about the size and branching of a molecule and larger molecules are expected to have a higher number of rings.

The second descriptor in the best two-variable model was *nHDon*. This is a simple count descriptor that accounts for the number of hydrogen bond donors. These are defined by the software as hydrogen atoms that are bound to a nitrogen or an oxygen atom.[32]

The coefficients [$\beta_i$ in Equation (1)] of both descriptors in the model show a negative sign, while the descriptor values ($x_i$) are by definition nonnegative. This gives an overall negative product. A negative value in the logistic regression equation pushes the final probability value towards zero. Solvate formation is more likely when the probability (*x*) is closer to zero, so the negative coefficients indicate that the higher the value of these two descriptors are, the more likely solvate formation is.

The average *MSE* of the model that uses *nCIC* and *nHDon* over 10 samples was 0.157, which is close to 0.148; the average *MSE* of the original model (Table 3). The simpler model also showed the same robust behavior over 10 random samples as the original model.

***3.3.2 The methanol model:*** In the methanol dataset, the descriptors that gave the best predictive ability are *TRS* (Total Ring Size) and *nHDon*. *TRS* is the sum of the number of atoms in each independent ring in the molecule (e.g. *TRS* value of benzene is 6 and of naphthalene is 12). This descriptor can probably be influential due to the stabilization of the solvate structures by the hydrophobic interactions between the rings.

***3.3.3 The dichloromethane model:*** In dichloromethane, the first descriptor was *SM3_H2*. This descriptor refers to the third order spectral moment of the reciprocal squared distance matrix (H2)[43]. The third spectral moment is calculated as the trace of the third power of the matrix.[44] This descriptor also describes the size and branching of molecules. Here again, the *SM3_H2* descriptor is not an easy one to estimate by looking at the molecular graph. Fortunately, a simple path count descriptor (*MPC01*) showed to be very similar to *SM3_H2*, with a high correlation (r = 0.983). *MPC01* is the count of paths of length 1 in the H-depleted molecular graph. In other words, it is equal to the number of bonds between non-hydrogen atoms in the molecular graph.[45-46] Both spectral moments and path counts increase exponentially with the size of the molecule, so their values were subject to a logarithmic transformation [x' = ln (1+x)], i.e., *SM3_H2* is obtained by the logarithmic transformation of the spectral moment, and *MPC01* by the logarithmic transformation of the count described above.

The second descriptor was *Hy*, which is called the hydrophilic factor. This factor is calculated using the formula in Equation (5).

$$H_y = \frac{(1+N_{Hy})\cdot log_2(1+N_{Hy})+nC\cdot\left(\frac{1}{nSK}.log_2\frac{1}{nSK}\right)+\sqrt{\frac{N_{Hy}}{nSK^2}}}{log_2(1+nSK)} \qquad (5)$$

Where $N_{Hy}$ is the number of hydroxyl, amine or thiol groups, nC is the number of carbon atoms and nSK is the number of non-hydrogen atoms.[47] This descriptor is highly correlated (r > 0.95) to the number of hydrogen bond donors (*nHDon*) descriptor, which can be used instead and the resulting model still gives similar results. This alternative simple model had an average *MSE* of 0.150 over 10 samples compared to 0.146 for the original model.

***3.3.4 The chloroform model:*** The first descriptor in chloroform was the same as for dichloromethane, *i.e. SM3_H2*. Again, this descriptor has a correlation of 0.984 with *MPC01* in the chloroform dataset. *SM3_H2* was combined with *H-050* to give the best model. The descriptor *H-050* is the number of hydrogen atoms attached to a heteroatom.[48-49] Here again, chloroform behaves in a similar manner to dichloromethane, where they share the same first descriptor and have two second descriptors that are almost identical (r > 0.95 correlation). The Average *MSE* of the simpler model, again over 10 samples is 0.152, compared to 0.148 of the original model.

***3.3.5 The water model:*** The best two-variable model of hydrate formation utilizes the $\pi ID$ and the *Mor05u* descriptors. The descriptor $\pi ID$ is the logarithmic transform of the conventional bond order ID number.[50] It is calculated using the formula

$$\pi ID = \ln(1 + nSK + \textstyle\sum_p w_p), \qquad (6)$$

Where $nSK$ is the number of non-hydrogen atoms and $w_p$is the weight of molecular path $p$. The index $p$ runs over all possible bond paths in the hydrogen-depleted molecular graph from the length of 1 bond to the longest possible. The weight assigned to each path, $w_p$, is the product of the conventional bond orders of all bonds in the path. The conventional bond order of single bonds is 1, for aromatic bonds it is 1.5, for double bonds it is 2 and for triple bonds it is 3. The value of this descriptor is affected by the size and branching and the type of bonds in a molecule. This gives information not only about the complexity, but also about the rigidity of a molecule.

For a simpler description, the *nCIC* descriptor can be used, which is the number of rings in the molecule and it shows a strong correlation with $\pi ID$ (r =0.854).

The second variable in the model was *Mor05u*. This is one of the 3D-MoRSE (3D-Molecule Representation of Structures based on Electron diffraction) descriptors. The 3D-MoRSE descriptors are calculated from the atomic 3D coordinates obtained by a molecular transform that is analogous to electron diffraction formulae.[51]

$$Mor05u = \sum_{i=1}^{nAT-1} \sum_{j=i+1}^{nAT} \frac{\sin(5r_{ij})}{5r_{ij}} \tag{7}$$

Where $r_{ij}$ is the distance between atoms i and j in the molecule and $nAT$ is the number of atoms. This descriptor requires previous knowledge of the 3D coordinates of the molecules under study, which is not always suitable for prediction. In order to keep the models simple and preserve their ability to describe solvate formation using the 2D molecular graph only, it is possible to take a highly correlated, easy-to-calculate descriptor instead. The number of hydrogen atoms (*nH*) and the number of atoms of molecule (*nAT*) descriptors are both highly correlated (r =0.94) with *Mor05u*. A model utilizing the $\pi ID$ and $nH$ descriptors has an average *MSE* of 0.161 compared to an average *MSE* of 0.159 of the original hydrate model. A model fitted using $nCIC$ along with $nH$ has an average *MSE* of 0.165. A table of the simple alternative models and their performance is given in Table 3. A more detailed version of the table is given in the Supporting Information (Table S8). For better understanding of the descriptors, the values of all mentioned descriptors for two drug molecules are given in Table S4, Supporting Information.

**Table 3.** The average performance of the simplified two-variable models over 10 samples in the 5 solvents.

| Model | Descriptors | Intercept | Descriptor 1 coefficient | Descriptor 2 coefficient | No. of data points | *MSE* | *AIC* | *AUC* |
|---|---|---|---|---|---|---|---|---|
| Ethanol | *nCIC* + *nHDon* | 3.994 | -0.766 | -0.889 | 1377 | 0.157 | 1320 | 0.854 |
| Methanol | *TRS* + *nHDon* | 2.808 | -0.084 | -0.612 | 3035 | 0.180 | 3268 | 0.810 |
| Dichloromethane | *MPC01* + *nHDon* | 13.236 | -3.649 | -0.339 | 2592 | 0.150 | 2428 | 0.864 |
| Chloroform | *MPC01* + *H-050* | 12.416 | -3.416 | -0.358 | 2384 | 0.152 | 2254 | 0.861 |
| Water | *nCIC* + *nH* | 2.731 | -0.606 | -0.088 | 607 | 0.165 | 597 | 0.835 |

***3.3.6 General discussion of the models:*** Regardless of the exact descriptor that turned up to be the best in each solvent, all two-variable models utilized one descriptor that measures the size and branching of the molecules and another one that is related to the hydrogen atoms in the molecules. Having similar descriptors in all models does not mean that these models are identical. The intercepts and the coefficients of these descriptors vary widely between the models, as can be seen in Table 2.  The difference in coefficients can be illustrated using the ethanol and methanol models. Although they share the same second descriptor (*nHDon*), the relative importance of this descriptor in ethanol is almost 1.5 times higher than it is in methanol.

The addition of a second variable to the models showed a notable reduction of the *MSE* in both ethanol and methanol, while it did not show a major reduction of *MSE* in the rest of the solvents.

This shows that hydrogen bonding has a large effect on the formation of ethanol and methanol solvates.

Surprisingly, introduction of the hydrogen bond-related descriptors did not improve the predictive ability of the models for hydrate formation. This does not imply that hydrogen bonding is not important in hydrates, but shows that the information given by hydrogen bonding descriptors in this specific dataset is already represented by the size and branching-related descriptors. A model that is fitted using the number of hydrogen bond acceptors (*nHAcc*) alone gives an average *MSE* of 0.237. This shows that hydrogen bonding is an important factor in hydrate formation, but it is not the most important discriminating factor according to this dataset. Indeed, the fact that single crystals of all these molecules were successfully grown from aqueous solutions suggests that even the non-hydrate formers among them are relatively hydrophilic.

The findings in these models agree with the common expectation that having a large, branched and rigid molecule makes the packing to optimally fill the three-dimensional space more difficult. The poor packing of molecules in the crystal seem to help the solvent molecules to diffuse through the structure and form a solvate. An example of the difference between the distribution of a size and branching descriptor for the solvate and non-solvate forming groups is shown in Figure 3.

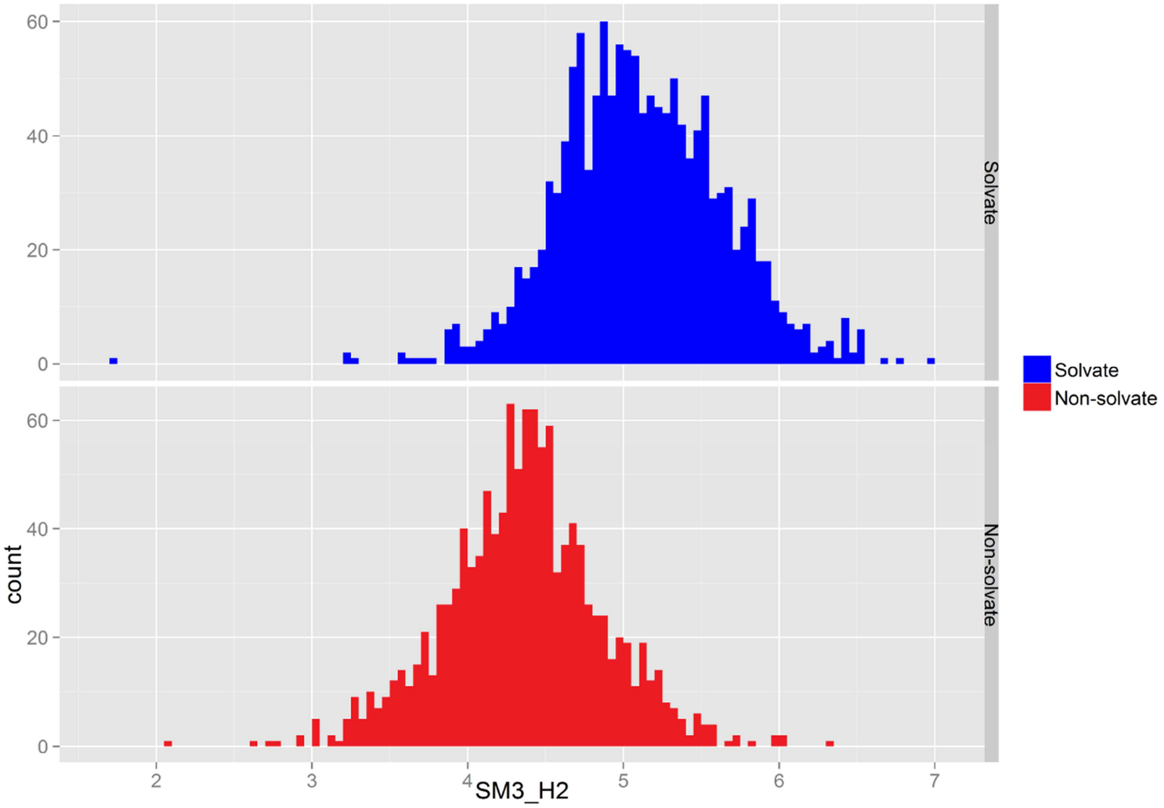


**Figure 3.** Histograms of the SM3_H2 descriptor distribution from the chloroform data

The availability of hydrogen bonds helps stabilizing the solvent in the voids of crystals, therefore giving a more stable hydrate or solvate. The importance of hydrogen bonding in solvate formation has been recognized in several publications.[52-53] The advancement in the current findings is the ability to quantify the relative importance of size, branching and hydrogen bonding. This aids the prediction of the ability of the molecules of interest to form a hydrate or a solvate relying only on the molecular structure.

## 3.4 Predicting the behavior of molecules

***3.4.1 Using the model equations:*** When a new molecule is to be predicted for solvate formation with one solvent, the values of the two descriptors in the model need to be calculated first. Afterwards, the descriptors values are fed into the logit function (Equation 1). An example of the use of the intercept and descriptor value to find $x$ in the simple hydrate model is given in Equation (8):

$$x = \frac{1}{1+e^{-(15.939\ -3.817AVS\_H2-0.861nHDon)}} \qquad (8)$$

The resulting value of *x* falls between 0 and 1.

***3.4.2 Is it a solvate or a non-solvate?*** In any binomial problem, there are only two possible outcomes. In the models above we consider a solvate to correspond with a predicted *x* value close to zero and a non-solvate to correspond with a value close to one. The cutoff point of the prediction, which tells whether our molecule of interest will form a solvate or not, should be close to 0.5. This is because equal size sampling was used for establishing these models. In order to select the optimum cutoff point in these models, the specificity (true positive predictions/all positive predictions) and sensitivity (true negative/all negative predictions) were used. The point that maximizes the specificity and the sensitivity was selected. In other words, the cutoff point can be chosen by finding the point where the specificity and sensitivity curves, plotted as the function of cutoff, cross. An automated script was developed in R to do this analysis. A representation of the cutoff point selection process can be seen in (Figure S6, Supporting Information). The cutoff point values for the five studied solvents were between (0.49-0.56).

This approach for cutoff point selection avoids the bias in the model towards one of the two groups (the solvate or the non-solvate group).

The final outcome of finding the cutoff point is the establishment of the decision boundary which splits the data into a solvate and non-solvate predicted region. Figure 4 shows how the classification system works for a sample of the dichloromethane dataset along with the decision boundary of the predictive model.

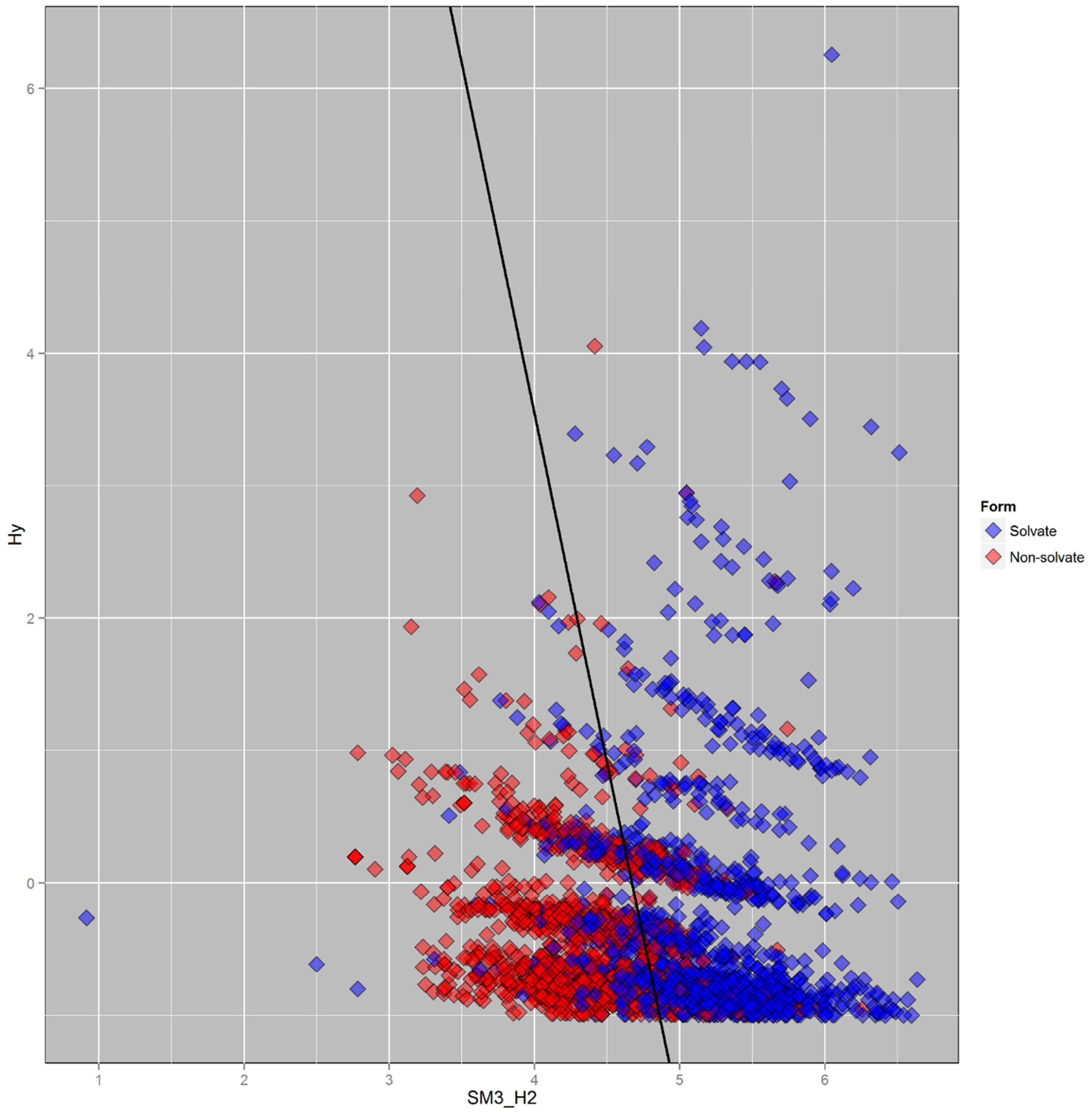
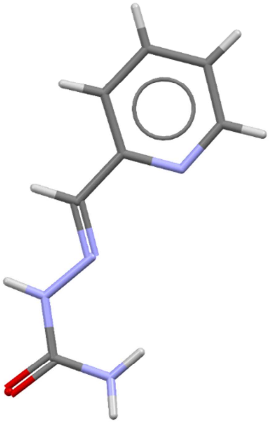


**Figure 4.** Distribution of the dichloromethane dataset using the best combination of two descriptors (2600 data points). The continuous line shows the optimized decision boundary, while the color of data points indicates the experimental solvate/non-solvate form of the corresponding molecule.

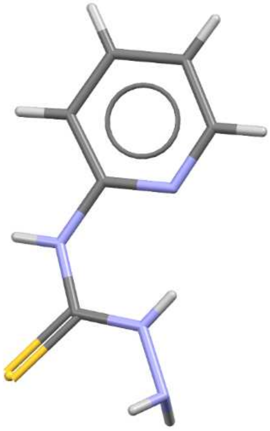
The complete dataset for each solvent was used to test the hydrate/solvate formation using the 2-variable models. The percentages of the correctly predicted data in different solvents were between 74 and 80%.

**3.5 Applicability of the models.** In this section, the factors that the models take into account and the factors that the models overlook are going to be discussed through examples from the datasets.

***3.5.1 Effects the models take into account.*** In this section, we illustrate the relative importance of the two main factors (size and branching, hydrogen bonding) using molecules from the ethanol dataset. In order to carry out this comparison, two molecules possessing different values of the two descriptors are going to be discussed. 1-((E)-2-pyridinylmethylidene)semicarbazone, CSD refcode: KUHGEA[54] and N-(pyridin-2-yl)hydrazinecarbothioamide, CSD refcode: XAPTOY[55] (Figure 5), are molecules that were both recrystallized from ethanol, but were not able to form ethanol solvates, despite the availability of multiple accessible hydrogen bond donor sites.



(a)

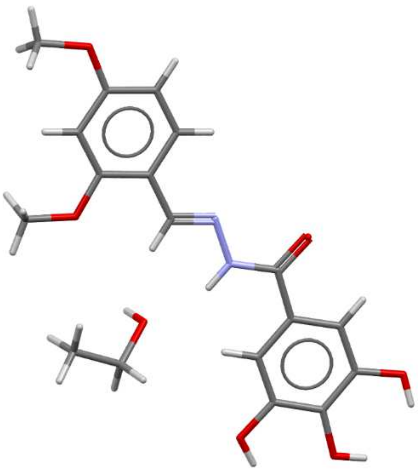


(b)

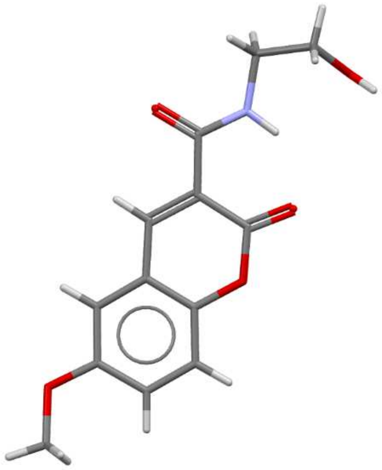
**Figure 5.** Molecular structures of (a) KUHGEA and (b) XAPTOY

The inability of these molecules to accommodate an ethanol molecule in their crystal structure was predicted correctly by the ethanol model. Even though these molecules have multiple hydrogen bond donors, their *AVS_H2* value is not high enough to surpass the decision boundary of the ethanol model (Figure S7, Supporting Information) into the solvate region. This proves that the effect of the size and branching of a molecule is more important than the number of hydrogen bond donors. The *AVS_H2* values for the KUHGEA and XAPTOY molecules are 2.972 and 2.979 and their *x* values according to the model were 0.882 and 0.879, respectively. This indicates that these two molecules have a low chance of forming an ethanol solvate. Further similar examples are given in Figure S8, while solvate entries with few donors, but larger, more branched structures are listed in Table S10 (Supporting Information).

Although the size and branching of molecules turned out to be the most important factor in determining solvate formation, the effect of hydrogen bonding cannot be ignored for alcohol solvates. The importance of the number of hydrogen bond donors, which was the second variable in the ethanol model, can be shown by comparing two molecules with similar *AVS_H2* values, but different number of hydrogen bond donors.



(a)



(b)

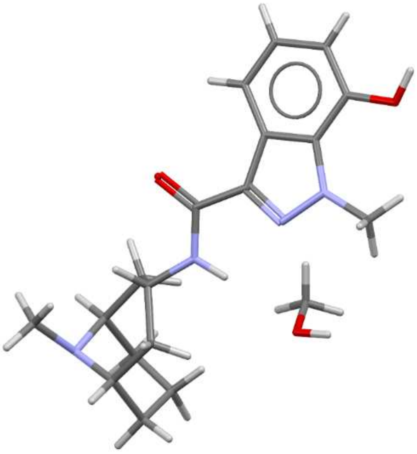
**Figure 6.** Molecular structures of (a) SOYQON and (b) UHUWEA.

N'-(2,4-Dimethoxybenzylidene)-3,4,5-trihydroxybenzohydrazide ethanol solvate, CSD refcode: SOYQON[56] (Figure 6a) was recrystallized from ethanol and formed an ethanol solvate. It has an *AVS_H2* value of 3.534 and 4 hydrogen bond donors. This structure was predicted correctly by the model to from the ethanol solvate ($x$ = 0.27). N-(2-hydroxyethyl)-6-methoxy-2-oxo-2H-chromene-3-carboxamide, CSD refcode: UHUWEA,[57] was also recrystallized from ethanol, but was not able to form an ethanol solvate (Figure 6b). It has an *AVS_H2* value of 3.513, but differs from the former molecule in that it has only 2 hydrogen bond donors. The inability of this molecule to form an ethanol solvate was also predicted correctly by the model ($x$ = 0.692). The effect of hydrogen bonding was accounted for in the models, hence the correct prediction of the behavior of these two molecules. Further examples are shown in the Supporting Information (Figures S9 and S10).

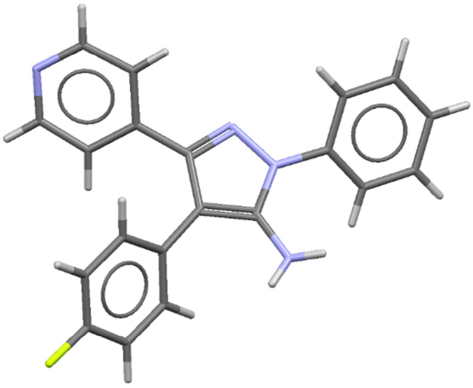
***3.5.2 Effects the models do not take into account.*** The examples shown so far are the clear-cut molecules, where the model worked excellently. For the molecules that were misclassified by the model, there must be some other factors that the model did not take into account. These factors were not identified by the models either because they do not show statistical significance (i.e., low number of examples) or for the fact that no descriptor looks at some of them. For these reasons, misclassified entries were surveyed manually and possible reasons were identified. The principal reasons found for misclassification, seen throughout the dataset will be discussed in this section.

*Hydrogen bonding strengths:* The current model takes into account only the number of hydrogen bond donors, but not their strength. The strength of a hydrogen bond depends on the nature of the functional groups in which the donor and acceptor atoms are located. This effect was studied thoroughly by different research groups,[58-59] who were able to represent the relative ability of several functional groups to donate or accept hydrogen bonds using empirical hydrogen-bond scales. It is important to mention that by taking into account the donor/acceptor coefficients tabulated by these research groups,[59] it was possible to explain a big part of the data misclassified by the models.

One case that clearly illustrates this is a comparison between 7-hydroxy-1-methyl-*N*-(9-methyl-9-azabicyclo[3.3.1]non-3-yl)-1H-indazole-3-carboxamide methanol solvate, CSD refcode: VUQMEA[60] and 4-(4-Fluorophenyl)-1-phenyl-3-(pyridin-4-yl)-1H-pyrazol-5-amine, CSD redcode: LANRUP.[61] Both compounds were crystallized from methanol and the values of the descriptors used by the methanol model, *TRS* and *nHDon* are identical for both (23 and 2, respectively). Their structures are shown in Figure 7.



(a)



(b)

**Figure 7.** Molecular structures of (a) VUQMEA and (b) LANRUP.

Although both these structures have two hydrogen bond donors each, the hydrogen bonding functional groups are not similar. While VUQMEA has an amide and a hydroxyl group, LANRUP shows one primary amine group. Amides are known to form stronger hydrogen bonds than amines.[62] To the model, both structures are identical, where they were both predicted to form a solvate with $x$ = 0.408, but in reality the strong amide donor of VUQMEA forms a hydrogen bond with the methanol molecule, and contributes to its retention in the crystal structure. The empirical donor coefficients mentioned earlier work well to explain these molecules. For example, Abraham[59] assigned a hydrogen bond donating constant of 0.4-0.55 to aliphatic amides and 0.08-0.16 to aliphatic amines. The hydrogen bond donating constant for the methanol hydroxyl group ranged between 0.31-0.37. On the other hand, the hydrogen bond accepting constants for the same functional group ranged between 0.48-0.6. This explains why the hydroxyl group of the methanol preferred to serve as a hydrogen bond acceptor in the VUQMEA structure case.  These are warning signs that the models, at the moment, cannot be used without considering the strength of the hydrogen bonding groups in the molecule and the solvent. The effect of different hydrogen bonding strengths on solvate formation was apparent in multiple cases. A number of paired solvate/non-solvate examples from the methanol dataset representing the case are listed in Table S11. Additional illustrations are shown in the Supporting Information (Figure S11 and Figure S12).

*Halogen bonding:* As can be expected, this type of interaction can be easily observed among the chlorinated solvents (dichloromethane and chloroform). Over 25 % of the solvate entries in the chloroform dataset show a short contact (at least 0.1 Å shorter than the sum of the van der Waals radii) between the chlorine atom of the solvent and a halogen bond acceptor (N, O, F, S, Br, or I) in the molecule. This indicates that halogen bonding is one of the main stabilizing interactions for these solvents in the crystal structure. One example of these structures is *t*-butyl (1-((4-bromophenyl)sulfonyl)-4-(4-methyl-1H-1,2,3-triazol-1-yl)piperidin-3-yl)carbamate chloroform solvate, CSD refcode: KUWWOP (Figure 8),[63] from the misclassified chloroform dataset. Based on the size, branching and the number of polar hydrogen atoms, this molecule was predicted not to form a chloroform solvate at a probability of 0.534. Although other contributors to the stabilization of chloroform in this crystal structure exist, it is notable that the chlorine atom of the chloroform is at a short distance (3.118 Å) from an $sp^2$ nitrogen atom (Figure 8). It appears reasonable that taking into account the possibility of forming a halogen bond could shift the predicted *x* value below the cutoff value (0.514).

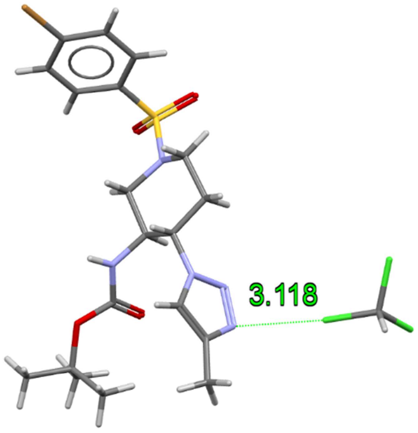
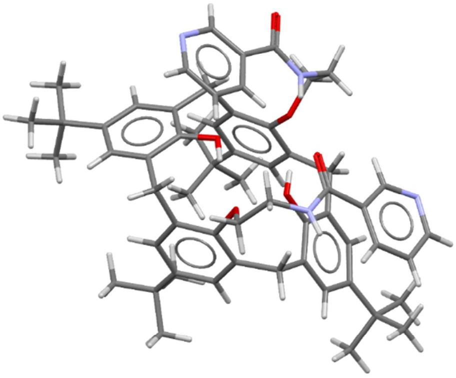


**Figure 8.** KUWWOP structure showing the distance between the chlorine of the chloroform and the nitrogen of the molecule (Å).

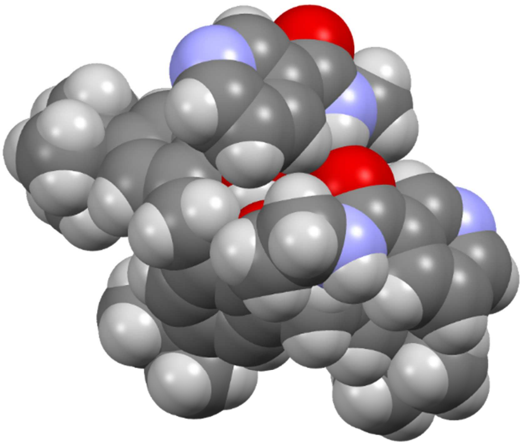
Halogen bonding does not seem to be sufficiently strong to retain the solvent in the crystals on its own, as no example of a solvate was found in the dataset where the only short contact between the solvent and the other molecule is a halogen bond. Nevertheless, this type of bond certainly contributes to the attractive interactions. Multiple examples of mispredictions involving a halogen bond were observed in the chloroform dataset. The reference codes of some of these cases are shown in Table S12. An illustration of the same case with dichloromethane is also shown in Figure S13.

Quantum chemical calculations and electrostatic potential approaches[64-65] suggest that the energy of the halogen-nitrogen bond can reach up to 29 kJ/mol with iodine and ca. 10.5 kJ/mol with chlorine at a distance of 2.9 Å.

*Availability of functional groups:* This factor becomes important when groups that affect hydrate and solvate formation (such as hydrogen-bonding groups) are present in the molecule, but they are not accessible by the solvent. One example showing the effect of low accessible surface area is 5,11,17,23-tetra-*t*-butyl-25,27-*bis*(2-(N-(pyrid-3-ylcarbonyl)amino)ethoxy)-26,28-dihydroxycalix[4]arene, CSD refcode: AZOMIL[66] from the methanol dataset.



(a)



(b)

**Figure 9.** (a) Capped-stick and (b) space filling representation of the AZOMIL molecule

By looking at the capped sticks model, it can be seen that this structure possesses four hydrogen bond donors. Three of these donors are involved in intramolecular hydrogen bonding, which renders them unavailable for intermolecular bonding according to Etter's rules.[67] By comparing the capped sticks to the space filling representation of the molecule shown in Figure 9, it can be noticed that the accessibility of these hydrogen bond donors as well as the remaining hydrogen bond donor is low due to steric effects, hence the inability of this molecule to form a solvate. This molecule was predicted by the model with a high probability to form a methanol solvate ($x$ = 0.017). A large part of this misprediction can be attributed to the inability of the model to estimate the accessibility of the hydrogen bond donors. An additional illustration and reference codes of examples that were predicted incorrectly for the same reason are provided in the Supporting Information (Figure S14, Table S13).

## 4. Conclusions

We have demonstrated that the use of molecular descriptors and machine learning techniques can identify molecular properties that contribute to solvate formation and can yield models with good predictive power. The size and branching of molecules was found to be the most important factor in each of the five solvents studied, while the presence of hydrogen bonding groups came second. Large, branched molecules optimize their close packing through interspersed solvent molecules,[68-70] while hydrogen bonding groups are responsible for inclusion of solvents via strong, specific interactions.[68,71] The present work expands on earlier research by quantifying the relative importance of these effects. While the broad factors are the same for each solvent, their weights are different. This means that the same molecule will be predicted to form a solvate with a different likelihood for each solvent.

The five models, one for each solvent, were able to correctly predict whether the molecules form a solvate for 74-80% of a 19,010 organic molecule dataset. The models are easy to use and provide instant predictions based on a minimal amount of information, *i.e.* the chemical formula. These attributes make the method well suited for a quick selection of suitable solvents when detailed experimental screening is not feasible, such as before the first recrystallization of a newly synthesized compound or when planning early pre-formulation tests of a drug candidate.

Analysis of incorrectly predicted results highlighted some limitations, which are mostly related to somewhat simplistic description of specific interactions. The descriptor set we used includes only simple counts of hydrogen bond donors and acceptors, but neither their strengths nor their steric accessibility are accounted for. It is expected that by devising appropriate descriptors for these effects, models with a higher success rate could be developed. The same applies for halogen bonds and competing intramolecular interactions.

While the present work is sufficient to demonstrate the potential of simple, descriptor based models, we plan further work to make the approach more generally applicable. First, we intend to survey a wider range of solvents, e.g. by including polar aprotic ones. Secondly, we wish to link the independently fitted solvent-specific models to each other through targeted screening experiments, which will involve screening the same set of drug molecules in each solvent. This is important to ensure that there is no relative bias between the different models, i.e. that the same predicted *x* value corresponds to the same experimental likelihood of solvate formation in each model. Such a set of linked models would provide an ideal tool for selecting solvents that are least likely to lead to unexpected solvate formation.

## ASSOCIATED CONTENT

**Supporting Information**. Explanation of the reciprocal squared topological distance matrix, sample descriptor values and further details of the models.

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

## AUTHOR INFORMATION

### Corresponding Author

*E-mail: L.Fabian@uea.ac.uk. Telephone: + 44 1603 591091

### Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

### Funding Sources

University of East Anglia.

ACKNOWLEDGMENT

K.T. is grateful for the financial support of the University of East Anglia through a PhD scholarship. Part of the analysis presented in this paper was carried out on the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia. The help and support of colleagues in the Drug Delivery and Pharmaceutical Materials group at the University of East Anglia is gratefully acknowledged.

REFERENCES

1. Grant, D. J. W.; York, P., *Int. J. Pharm.* **1986**, *30*, 161-180.
2. Clarke, H. D.; Arora, K. K.; Bass, H.; Kavuru, P.; Ong, T. T.; Pujari, T.; Wojtas, L.; Zaworotko, M. J., *Cryst. Growth Des.* **2010**, *10*, 2152-2167.
3. Vippagunta, S. R.; Brittain, H. G.; Grant, D. J. W., *Adv. Drug Deliv. Rev.* **2001**, *48*, 3-26.
4. Stephenson, G. A.; Groleau, E. G.; Kleemann, R. L.; Xu, W.; Rigsbee, D. R., *J. Pharm. Sci.* **1998**, *87*, 536-542.
5. Khankari, R. K.; Law, D.; Grant, D. J. W., *Int. J. Pharm.* **1992,** *82*, 117-127.
6. Peterson, M. L.; Hickey, M. B.; Zaworotko, M. J.; Almarsson, Ö., *J. Pharm. Pharmaceut. Sci.* **2006**, *9*, 317-326. **Check:**
7. Tian, F.; Qu, H.; Zimmermann, A.; Munk, T.; Jørgensen, A. C.; Rantanen, J., *J. Pharm. Pharmacol.* **2010**, *62*, 1534-1546.
8. Braun, D. E.; Karamertzanis, P. G.; Price, S. L., *Chem. Commun.* **2011**, *47*, 5443-5445.
9. Dunitz, J. D., *Chem. Commun.* **2003**, 545-548.
10. Morissette, S. L.; Almarsson, Ö.; Peterson, M. L.; Remenar, J. F.; Read, M. J.; Lemmo, A. V.; Ellis, S.; Cima, M. J.; Gardner, C. R., *Adv. Drug. Deliv. Rev.* **2004**, *56*, 275-300.
11. Bardwell, D. A.; Adjiman, C. S.; Arnautova, Y. A.; Bartashevich, E.; Boerrigter, S. X. M.; Braun, D. E.; Cruz-Cabeza, A. J.; Day, G. M.; Della Valle, R. G.; Desiraju, G. R.; van Eijck, B. P.; Facelli, J. C.; Ferraro, M. B.; Grillo, D.; Habgood, M.; Hofmann, D. W. M.; Hofmann, F.; Jose, K. V. J.; Karamertzanis, P. G.; Kazantsev, A. V.; Kendrick, J.; Kuleshova, L. N.; Leusen, F. J. J.; Maleev, A. V.; Misquitta, A. J.; Mohamed, S.; Needs, R. J.; Neumann, M. A.; Nikylov, D.; Orendt, A. M.; Pal, R.; Pantelides, C. C.; Pickard, C. J.; Price, L. S.; Price, S. L.; Scheraga, H. A.; van de Streek, J.; Thakur, T. S.; Tiwari, S.; Venuti, E.; Zhitkov, I. K., *Acta Crystallogr., Sect. B* **2011**, *67*, 535-551.
12. Pino-Mejías, R.; Cubiles-de-la-Vega, M. D.; Anaya-Romero, M.; Pascual-Acosta, A.; Jordán-López, A.; Bellinfante-Crocci, N., *Environ, Modell. Softw.* **2010**, *25*, 826-836.
13. Bellazzi, R.; Zupan, B., *Int. J. Med. Inf.* **2008**, *77*, 81-97.

14. Krilavičius, T.; Morkevičius, V., *INTELLECTUAL ECONOMICS* **2011**, *5*, 224–243.

15. Allen, F. H., *Acta Crystallogr., Sect. B* **2002**, *B58*, 380-388.

16. Hofmann, D. W. M.; Apostolakis, J., *J. Mol. Struct.* **2003**, *647*, 17-39.

17. Infantes, L.; Fábián, L.; Motherwell, W. D. S., *CrystEngComm* **2007**, *9*, 65-71.

18. Nangia, A.; Desiraju, G.R., *Chem. Commun.* **1999**, 605-606.

19. van de Streek, J.; Motherwell, S., *CrystEngComm* **2007**, *9*, 55-64.

20. Brychczynska, M.; Davey, R. J.; Pidcock, E., *New J. Chem.* **2008**, *32*, 1754-1760.

21. Weiss, S. M.; Indurkhya, N., *Predictive data mining: a practical guide*. Morgan Kaufmann Publishers Inc.: San Francisco,1998; p 227.

22. Cheng, J.; Tegge, A. N.; Baldi, P., *IEEE Rev. Biomed. Eng.* **2008**, *1*, 41-49.

23. Fan, C.; Xiao, F.; Wang, S., *Appl. Energy* **2014**, *127*, 1-10.

24. Li, C.-S.; Chen, M.-C., *Neurocomputing* **2014**, *133*, 74-83.

25. Wicker, J.; Cooper, R., *CrystEngComm* **2015**, *17*, 1927-1934.

26. Johnston, A.; Johnston, B. F.; Kennedy, A. R.; Florence, A. J., *CrystEngComm* **2008**, *10*, 23-25.

27. Liaw, A.; Wiener, M., *R News* **2002**, *2*, 18-22.

28. Görbitz, C. H.; Hersleth, H. P., *Acta Crystallogr., Sect. B* **2000**, *56*, 526-34.

29. Desiraju, G. R., *J. Chem. Soc., Chem. Commun.* **1991**, 426-428.

30. Todeschini, R.; Consonni, V., *Handbook of Molecular Descriptors*. Wiley-VCH: 2009.

31. Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. *Acta Crystallogr., Sect. B* **2002**,*58*, 389-397.

32. Talete srl, *DRAGON : Software for Molecular Descriptor Calculation*, 6.0; 2013.

33. R Core Team (2015). R: A language and environment for statistical
computing. R Foundation for Statistical Computing, Vienna, Austria.
URL http://www.R-project.org/.

34. Lowry, R., *Concepts and Applications of Inferential Statistics*. 2013. http://vassarstats.net/textbook/

35. Fisher, R. A., *Statistical methods for research workers*, 13th ed.; Oliver and Boyd: London, 1958, p. 336.

36. Hanrahan, G., *Artificial neural networks in biological and environmental analysis*, 1st ed.; CRC Press: Boca Raton, 2011, p 214.

37. Cortes, C.; Vapnik, V., *Mach. Learn.* **1995**, *20*, 273-297.

38. Kleinbaum, D.G.; Klein, M. *Logistic Regression. A Self-Learning Text*, 3rd ed.; Springer Science+Business Media: New York, 2010, p 702.

39. Hanley, J. A.; McNeil, B. J., *Radiology* **1982**, *143*, 29-36.

40. Akaike, H., *Psychometrika* **1987**, *52*, 317-332.

41. Akaike, H., *IEEE Trans. Autom. Contr.* **1974**, *19*, 716-723.

42. Stone, M., *J.R. Stat. Soc. Series B Stat. Methodol.* **1977**, *39*, 44-47.

43. Mihalić, Z.; Trinajstić, N., *J. Chem. Educ.* **1992**, *69*, 701.

44. Estrada, E., *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 844-849.

45. Randić, *M., MATCH Commun. Math. Comput. Chem.* **1979**, *7*, 5-64

46. Randić, M.; Brissey, G. M.; Spencer, R. B.; Wilkins, C. L., *Comput. Chem.* **1979**, *3*, 5-13.

47. Todeschini, R.; Vighi, M.; Finizio, A.; Gramatica, P., *SAR QSAR Environ. Res.* **1997**, *7*, 173-193.

48. Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K., *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163-172.

49. Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J., *J. Phys. Chem. A* **1998**, *102*, 3762-3772.

50. Randić, M.; Jurs, P. C., *Mol. Inform.* **1989**, *8*, 39-48.

51. Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V., *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1030-1037.

52. Jetti, R. K. R.; Griesser, U. J.; Krivovichev, S.; Kahlenberg, V.; Bläser, D.; Boese, R., *Acta Crystallogr., Sect. A* **2005**, *61*, c286.

53. Harmon, K. M.; Webb, A. C., *J. Mol. Struct.* **1999**, *508*, 119-128.

54. Garbelini, E. R.; Hörner, M.; Giglio, V. F.; da Silva, A. H.; Barison, A.; Nunes, F. S., *Z. Anorg. Allg. Chem.* **2009**, *635*, 1236-1241.

55. Rapheal, P. F.; Manoj, E.; Kurup, M. R. P.; Suresh, E., *Acta Crystallogr., Sect. E* **2005**, *61*, o2243-o2245.

56. Alhadi, A. A.; Saharin, S. M.; Mohd Ali, H.; Robinson, W. T.; Abdulla, M. A., *Acta Crystallogr., Sect. E* **2009**, *65*, o1373.

57. Santos-Contreras, R. J.; Martínez-Martínez, F. J.; Mancilla-Margalli, N. A.; Peraza-Campos, A. L.; Morín-Sánchez, L. M.; García-Báez, E. V.; Padilla-Martínez, I. I., *CrystEngComm* **2009**, *11*, 1451-1461.

58. Hunter, C. A., *Angew. Chem. Int. Ed. Engl.* **2004**, *43*, 5310-5324.

59. Abraham, M. H.; Platts, J. A., *J. Org. Chem.* **2001**, *66*, 3484-3491.

60. Vernekar, S. K. V.; Hallaq, H. Y.; Clarkson, G.; Thompson, A. J.; Silvestri, L.; Lummis, S. C. R.; Lochner, M., *J. Med. Chem.* **2010**, *53*, 2324-2328.

61. Abu Thaher, B.; Koch, P.; Schollmeyer, D.; Laufer, S., *Acta Crystallogr., Sect. E* **2012**, *68*, o632.

62. McMurry, J. E.; Hoeger, C. A.; Peterson, V. E.; Ballantine, D. S., *Fundamentals of General, Organic, and Biological Chemistry: Pearson New International Edition*, 7th ed. Pearson: Harlow, 2013, p 976.

63. Schramm, H.; Saak, W.; Hoenke, C.; Christoffers, J., *Eur. J. Org. Chem.* **2010**, *2010*, 1745-1753.

64. Politzer, P.; Lane, P.; Concha, M. C.; Ma, Y.; Murray, J. S., *J. Mol. Model.* **2007**, *13*, 305-311.

65. Valerio, G.; Raos, G.; Meille, S. V.; Metrangolo, P.; Resnati, G., *J. Phys. Chem. A* **2000**, *104*, 1617-1620.

66. Yu, L.; Hao, W.; Heng-Yi, Z.; Bang-Tun, Z.; Li-Hua, W., *J. Supramol. Chem.* **2002**, *2*, 515-519.

67. Etter, M. C., *Acc. Chem. Res.* **1990**, *23*, 120-126.

68. Price, C. P.; Glick, G. D.; Matzger, A. J., *Angew. Chemie - Int. Ed.* **2006**, *45*, 2062–2066.

69. Roy, S.; Quiñones, R.; Matzger, A. J., *Cryst. Growth Des.* **2012**, *12*, 2122–2126.

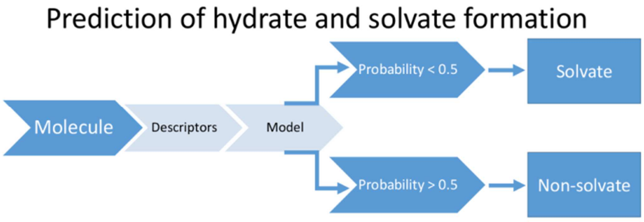
70. Bērziņš, A.; Skarbulis, E.; Rekis, T.; Actiņš, A., *Cryst. Growth Des.* **2014,** *14*, 2654–2664.

71. Coutinho, K.; Cabral, B. J. C.; Canuto, S., *Chem. Phys. Lett.* **2004**, *399*, 534–538.

## For Table of Contents Use Only

# Prediction of hydrate and solvate formation using statistical models

*Khaled Takieddin, Yaroslav Z. Khimyak, and László Fábián**



Models were developed to predict the likelihood of solvate formation by neutral organic molecules with methanol, ethanol, chloroform, dichloromethane and water. Only the structural formula of the molecules is required as input.