molecular pharmaceutics

Article

pubs.acs.org/molecularpharmaceutics

In Silico Prediction of Aqueous Solubility: A Multimodel Protocol Based on Chemical Similarity

Florent Chevillard,[†,‡] David Lagorce,[†] Christelle Reynès,[†,§] Bruno O. Villoutreix,[†] Philippe Vayer,*[,∥] and Maria A. Miteva*[,†]

[†]Université Paris Diderot, Sorbonne Paris Cité, Molécules Thérapeutiques *in silico*, Inserm UMR-S 973, 35 rue Helene Brion, 75013 Paris, France

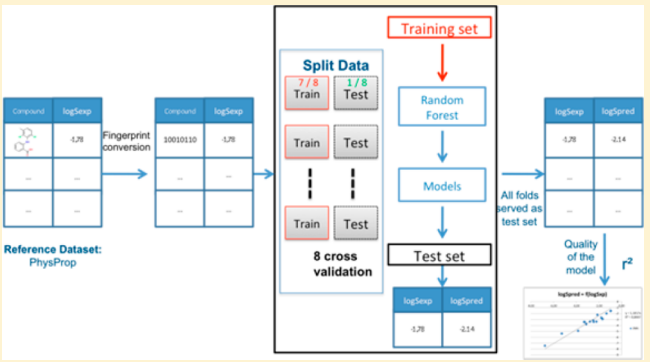
[‡]Institute of Pharmaceutical Chemistry, Phillips University Marburg, Marbacher Weg 6-10, 35037 Marburg, Germany

[§]Lab. Physique Industrielle et Traitement de l'Information EA 2415, UFR Pharmacie - Univ. Montpellier 1, 15 avenue Charles Flahault - BP 14491, 34093 Montpellier Cedex 5, France

[∥]BioInformatic Modelling Department, Technologie Servier, 45007 Orléans cedex1, France

Supporting Information

ABSTRACT: Aqueous solubility is one of the most important ADMET properties to assess and to optimize during the drug discovery process. At present, accurate prediction of solubility remains very challenging and there is an important need of independent benchmarking of the existing in silico models such as to suggest solutions for their improvement. In this study, we developed a new protocol for improved solubility prediction by combining several existing models available in commercial or free software packages. We first performed an evaluation of ten in silico models for aqueous solubility prediction on several data sets in order to assess the reliability of the methods, and we proposed a new diverse data set of 150 molecules as relevant test set, SolDiv150. We developed a random forest protocol to evaluate the performance of different fingerprints for aqueous solubility prediction based on molecular structure similarity. Our protocol, called a "multimodel protocol", allows selecting the most accurate model for a compound of interest among the employed models or software packages, achieving r^2 of 0.84 when applied to SolDiv150. We also found that all models assessed here performed better on druglike molecules than on real drugs, thus additional improvement is needed in this direction. Overall, our approach enlarges the applicability domain as demonstrated by the more accurate results for solubility prediction obtained using our protocol in comparison to using individual models.



KEYWORDS: *solubility prediction, chemical structure similarity, QSPR models, multimodel optimization*

INTRODUCTION

Aqueous solubility is one of the most important ADMET (absorption, distribution, metabolism, excretion and toxicity) properties to be optimized during the drug discovery process.[1,2] Poor solubility has been identified as the cause of many drug development failures,[3] and improving the aqueous solubility of bioactive molecules is a major issue in medicinal chemistry.[1−4] The rate of dissolution and permeability of drugs[3] strongly depend on its solubility. In addition, high concentrations of poorly soluble drugs in the human body may result in crystallization and toxicity. Recently, it has been found that real drugs are much more soluble[5] than those druglike molecules in the ZINC database passing Lipinski's rule of five.[1] Considering the critical role of solubility, its evaluation is crucial in all drug discovery projects. However, measuring experimentally the solubility for thousands and millions of molecules used in high throughput screening (HTS) is unrealistic[6] and impossible for millions of virtual molecules not yet synthesized but of potential interest. Therefore, prediction of solubility by in silico approaches would be highly valuable as it will assist the design and the prioritization of small molecules during the first steps of the drug discovery process but should also be beneficial to other commercially important compounds such as agrochemicals.

The thermodynamic solubility, denoted as S in moles per liter, is the maximum amount of the most stable crystal form of a compound that can remain in solution under thermodynamic equilibrium between the solid and dissolved state at a given temperature.[2] Considering the un-ionized form of the molecule, the thermodynamic solubility is stated as intrinsic solubility. Generally, thermodynamic solubility tends to be lower than

Received: April 26, 2012
Revised: October 9, 2012
Accepted: October 16, 2012

dx.doi.org/10.1021/mp300234q | *Mol. Pharmaceutics* XXXX, XXX, XXX−XXX

kinetic solubility, the latter depending on the compound crystal form or polyforms. The kinetic solubility is typically measured in a stock solution of the compound in dimethyl sulfoxide (DMSO), from which a sequential dilution is done in water.[2] Allowing sufficient time, the final form will be the most stable crystal form, and the solubility will approach the true thermodynamic solubility. The in silico predictions address mainly the thermodynamic solubility.

Accurate prediction of the solubility is a tremendous challenge for a large number of compounds since intermolecular adhesive interactions between solute−solute, solute−solvent, and solvent−solvent molecules involved in the dissolution process should be evaluated (those could be predicted by the first principle methods).[7] Thus, many different in silico approximations have been developed aiming at fast and, in some cases, accurate estimation of aqueous solubility of chemicals.[8−12] Historically, the first models were based on experimentally measured boiling/melting points, pK_a and the octanol−water partition coefficient (logP) values of the compounds. The classic way was indeed to combine the melting point with logP of the un-ionized molecule using the general solubility equation (GSE).[13−18] This technique cannot be applied to salt or lyophilized forms. Given that such experimental data are available for a limited number of compounds, other approaches are being developed, e.g., by using predicted logP.[7] Today the most commonly used methods are based on quantitative structure property relationship (QSPR) models allowing to correlate the aqueous solubility with various molecular descriptors (physicochemical, topological, 2D or 3D) using mathematical models.[5,19−21] Recent analyses stressed that sometimes the relationship between computed descriptors and the solubility is not straightforward[22] and that the applicability domain has to be considered.[23] Further, QSPR models need a high quality of experimentally measured solubility for the training sets, while it should be borne in mind that it has been estimated that the average error on the experimental values of aqueous solubility is probably more than 0.6 log unit for organic compounds.[24] Altogether, the availability of proper experimental solubility data, the applicability domain, as well as imperfections of the employed in silico techniques, demonstrate the need of independent benchmarking of the existing models and their improvement. Along these lines, a competition for accurate prediction of intrinsic solubility of 32 diverse druglike molecules with uniformly measured data, as in the proposed training set of 100 compounds, has recently been organized, "the Solubility Challenge".[8−10]

In this work, we developed a protocol that should improve in silico solubility prediction. First, we performed an evaluation of ten in silico models for aqueous solubility prediction on several data sets in order to assess the reliability of the methods. Then, we developed a random forest protocol with the goal to evaluate the performance of different fingerprints for the aqueous solubility prediction based on molecular structure similarity. Finally, we suggested and validated a new protocol, called a "multimodel protocol", combining several existing models available in commercial or free software, which allowed finding and selecting the most accurate model for a compound of interest among all the available models or software packages.

EXPERIMENTAL SECTION

Data Sets' Preparation. Four different data sets with available experimental values of intrinsic solubility (expressed in molar units mol/L[10]) were selected for solubility models' evaluation. We used FAF-Drugs2[25] to remove duplicate molecules (using canonical smiles), salts and inorganic compounds. We performed also filtering with FAF-Drugs2 for some physicochemical properties since solubility models have often been trained on druglike molecules[5,26] (MW < 500, hydrogen bond donors (HBD) < 5, hydrogen bond acceptors (HBA) < 10, number of heavy atoms < 37, −4 < logP < 5, rotatables bonds < 15; toxic/reactive groups were not removed). The *Standardize* protocol in Pipeline Pilot (www.accelrys.com/products/pipeline-pilot) with the parameter *NeutralizeBondedZwitterion* was used for the neutralization of molecules since the models assessed here predict intrinsic solubility that requires the neutral form of molecules.[5] In addition, the solubility software (except Pipeline) assessed here employ an internal standardization of the compounds. For the QikProp and VolSurf+ models using 3D descriptors, the 3D structures of the compounds were generated with Corina 3.4 (http://www.molecular-networks.com/products/corina). Only the lowest energy conformation among the 20 generated ones was kept for each molecule. In order to have different molecules in each data set the duplicates were removed as explained below.

Solubility Challenge Test Set. The Solubility Challenge data set is divided into two parts: Solubility Challenge training set of 100 compounds and Solubility Challenge test set of 32 compounds recently published.[8−10] The main strength of this data set is that it contains very diverse compounds with uniform experimental data of solubility. The intrinsic solubility was calculated using the CheqSol approach.[27] Among the 32 compounds, six were removed because reliable experimental values were not available. This procedure yielded 26 compounds.

PhysProp. 6152 compounds were extracted from the commercial PhysProp data set.[28] After applying the FAF-Drugs2 protocol (explained above) and removing duplicates with the Solubility Challenge test set, the number of compounds was reduced to 3970.

Solubility Challenge Training Set. The Solubility Challenge training set was prepared in the same way that the Solubility Challenge test set, but the duplicates with the PhysProp data set were removed, reducing the number of compounds to 71.

Huuskonen. This data set has been selected from the two commercial databases AQUASOL[14,29] and PhysProp,[28] and has been used as training set for creation of many models.[30,31] After applying the FAF-Drugs2 protocol and removing duplicates with all the other data sets, the number was decreased to 830.

Keys To Select the Most Appropriate Solubility Model. To determine the best fingerprint for the solubility prediction, we created models based on different molecular fingerprints using a nonparametric regression method for aqueous solubility prediction.

Random Forest Algorithm (RF). For the regression method, the Random Forest algorithm[32] was chosen. The method is based on an ensemble of decision trees, from which the prediction of a continuous variable, in this case the aqueous solubility, is obtained as the average of the predicted values of all trees. Each tree is an unpruned usual regression tree built on a subset of features and observations. Indeed, as the regression tree is a deterministic method, in order to obtain different trees, perturbations are added to each tree by performing a double randomization on features and observations. Random Forests

were trained using the randomForest library in the statistical computing environment R.[33]

Optimization of the Parameters for RF. Several statistical parameters can be tuned in order to improve the learning in a Random Forest algorithm. In this study, the most two influential parameters were optimized: *ntree*, which is the number of trees used to compute the final average predicted value, and *mtry*, which is the number of variables randomly chosen to build each individual tree. Scripts from R[33] were used, and both parameters were simultaneously optimized by using a grid search. The following ranges were proceeded: *ntree*, from 100 to 1000 by steps of 100, and *mtry*, from 20 to 200 by steps of 10. Both parameters, *ntree* and *mtry*, were optimized for each tested fingerprint independently.

Quality Measures. *Criteria To Evaluate Solubility Models.* To compare the performance of the solubility prediction models (solubility unit used is log(mol/L)) four criteria were used:

r^2, the squared regression coefficient for the correlation between experimental and predicted values;

P1, the ratio of molecules with a predicted error within 0.5 log (mol/L);

P2, the ratio of molecules with a predicted error within 1 log (mol/L);

RMSE, the root mean squared error.

Molecular similarity Evaluation. Molecular similarity used for the multimodel protocol was evaluated using the MACCS key fingerprints. These fingerprints were developed for substructure searching or for entire structure comparison.[34] They code the presence or absence of 166 molecular substructures. As a measure of similarity between two structures, the Tanimoto coefficient was applied, defined by

$$\text{similarity} = \frac{N_{A\cap B}}{N_A + N_B - N_{A\cap B}}$$

where N_A and N_B are the numbers of bits in bitstrings A and B, respectively, and $N_{A\cap B}$ is the number of bits which are common between the two bitstrings A and B. The measure of similarity is between 0.0 and 1.0, where 1.0 indicates strict equivalence of the bitstrings. Using the MACCS keys, a similarity of 1.0 usually means that the structures are identical (or at least very closely related) apart from stereochemistry, which is not taken into account by the keys.

Mean Absolute Error. In the multimodel protocol, the mean absolute error (MAE) for a target compound A was estimated using the absolute prediction errors of three compounds B, C and D (eB, eC and eD) with known experimentally measured solubility, which are structurally similar to compound A, using the following equation:

$$\text{MAE} = \frac{|eB| + |eC| + |eD|}{3}$$

RESULTS AND DISCUSSION

Chemical Space of Data Sets Used for Solubility Model Evaluation. An important criterion for this study was the chemical diversity of the data sets. Our study is based on four data sets filtered for physicochemical properties important for druglikeness (for details see Experimental Section): Huuskonen[29] (number of compounds = 830), PhysProp (www.srcinc.com) (number of compounds = 3970), and the Solubility Challenge test (number of compounds = 26) and training sets (number of compounds = 71).[8−10]

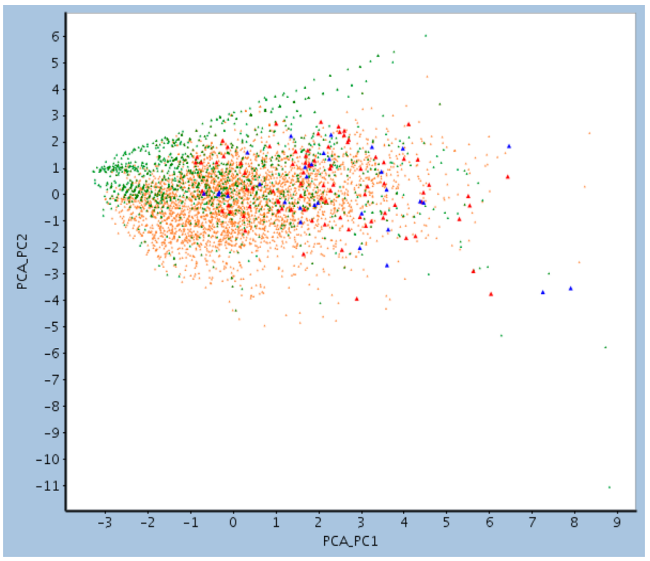


Figure 1. The first two PC scores for all data sets: The dots represent the Huuskonen data set (in green), the PhysProp (in orange), the Solubility Challenge training set (in red) and the Solubility Challenge test set (in blue).

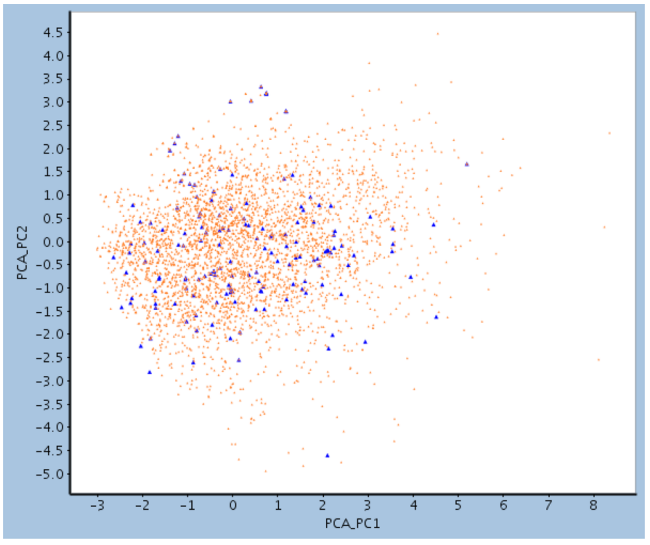


Figure 2. Representation of the 150 centroids selected for the test set SolDiv150 in the chemical space according to the two first PC scores. Blue dots represent the centroids; the orange dots represent the PhysProp data set.

In order to analyze the data sets' chemical space, we used a principal component analysis (PCA) learned on all compounds from the four data sets (4897 compounds) using the *Learn Molecular PCA model* available in Pipeline Pilot (www.accelrys.com). Nine properties of the compounds (molecular weight, logP, HBD, HBA, numbers of aromatic rings, rings, rotatable bonds, atoms and fragments) which are critical for solubility prediction[4,26,35] were used to compute the principal components (PC). The first two PC scores (with percentage of the variance 51 and 12) which explain best the global variability of the data (63%) were selected to be presented here. In this context, the compounds of each data set were then projected

dx.doi.org/10.1021/mp300234q | *Mol. Pharmaceutics* XXXX, XXX, XXX−XXX

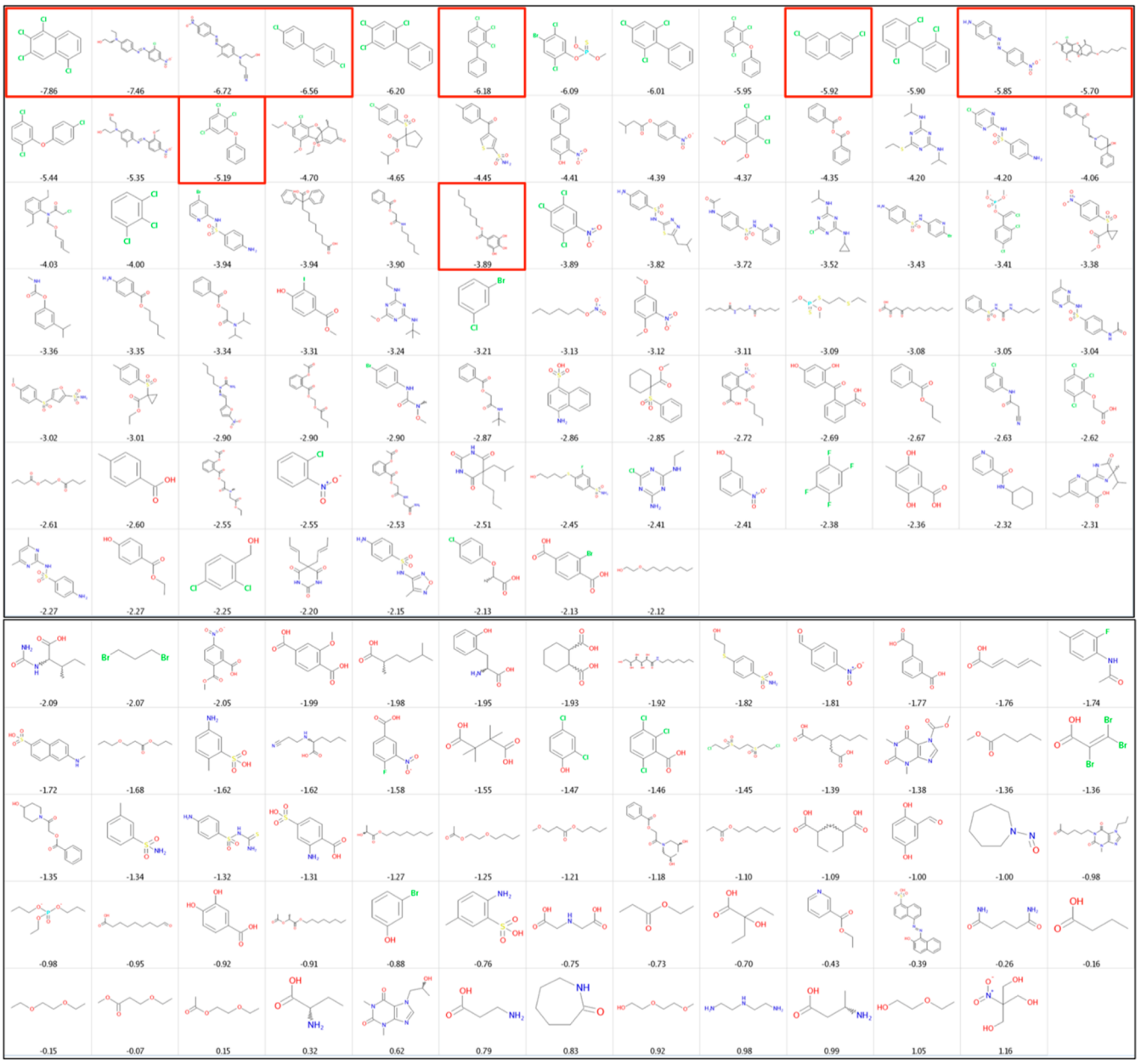


Figure 3. Molecular structures of the SolDiv150 data set with the experimental values of solubility taken from PhysProp. Ten molecules with an absolute error >2.0 log unit for the solubility logS as predicted by ISIDA are highlighted by red lines.

together onto this subspace in order to study their diversity (Figure 1).

In addition to the above-mentioned data sets, another diverse test set was prepared for solubility evaluation, called SolDiv150, that can also be useful for other benchmark studies. This data set contains 150 compounds from the PhysProp data set selected by clustering. The 150 centroids selected for the SolDiv150 set are shown in Figure 2, and their structures are given in Figure 3. These compounds were chosen because they represent the chemical space of PhysProp as illustrated in Figure 2, and as centroids of the obtained diverse clusters. The clustering protocol was performed with the MACCS key fingerprints (see Experimental Section for details) in Pipeline Pilot using a maximum distance of 0.3 (the Tanimoto coefficient) within the clusters. The centroid of each cluster was taken for the test set, if the cluster contained at least 4 compounds. As seen from Figure 3 the clustering approach allowed the selection of compounds from different chemical series.

Comparison of Existing Solubility Models. Ten QSPR models using 2D or 3D descriptors from commercial software or freeware available online were selected for evaluation (Table 1). QSPR models do not need any experimental data for the compound of interest but only the chemical structure of the compounds to predict its property (here solubility). However, the application of such a model is limited to the chemical space spanned by the compounds used to train the model, i.e., the applicability domain. QSPR models can be based on fragment-based approaches estimating the solubility by summing up the contribution of different fragments.[7,31,36,37] Yet, distinguishing isomers and/or missing fragments can be a problem for such methods.[5] Another type of model is based on molecular properties' similarity and requires various descriptor computations. All tools assessed here predict the intrinsic solubility and

dx.doi.org/10.1021/mp300234q | Mol. Pharmaceutics XXXX, XXX, XXX−XXX

Table 1. List of the Models Used in the Benchmark

models	descriptors	modeling method
Pipeline Pilot v.7.5: solubility[30]	electrotopological indices	artificial neural network
Pipeline Pilot v.7.5: ADMET-solubility[26]	atomic, topological	genetic algorithm, multiple linear regression
MOE v.2010.10[31,38]	fragments: 76 atom types	fragment-based, multiple linear regression
ACD lab v.12.0[39]	atomic	experiment-based combined with QSPR
QikProp v.3.4[40]	2D and 3D	experiment-based combined with QSPR
ADMET Predictor v.6[41]	atomic	3 components partial least squares
volsurf+ 1.0.6[42,43]	3D	QSPR combined with 3D descriptors
FAF-Drugs2[25,35]	atomic, topological	multiple linear regression
ALOGpS v.2.1[44] – VCC lab	electrotopological indices	artificial neural network
ISIDA[45]	fragments	QSPR, clustering

provide values in molar units (mol/L). The results obtained independently on the 5 different data sets are shown for each data set in Table 2.

According to the results, it is important to underline the fact that there is no model that truly outperforms the others. For instance, ACD, ADMET Predictor and VCC lab perform well on the PhysProp data set, but perform poorly on the Solubility Challenge sets. Similarly, the two PP models and FAF do not perform satisfactory on both Solubility Challenge sets. In fact, most of the software do not perform very well on the Solubility Challenge test set, stressing that the current models will have to be improved. If we focus on the SolDiv150 set, we note that the solubility values were poorly predicted by ISIDA (P1 = 0.33). The worst predictions for ISIDA with an absolute error for logS > 2.0 were obtained for 10 molecules underlined in red in Figure 3. In this case, all compounds were predicted to be more soluble than the experimental data. We can distinguish 3 common fragments for these molecules shown in Figure 4 that may have been insufficiently represented when the model was trained. Another reason for the first two fragments could be planarity that may not be sufficiently taken into consideration in the model, since usually the planarity of the molecules can lead to a decreased solubility.[4]

Overall, only the results on the Huuskonen's data set are satisfactory for all models, possibly because many models were trained on these data. This observation outlines the difficulty to test models with public data because often the test sets contain compounds similar to those of the training sets and the need to develop new freely available and clean, experimentally measured, solubility data sets for benchmarking.

Since the performance of the assessed models significantly varies depending on the test sets used, we decided to combine all models in an attempt to improve the accuracy of the prediction. A combination of these models could increase the effectiveness of the prediction for new compounds by enlarging the applicability domain due to the different training sets used initially to develop the different models

Multimodel Protocol. The above presented analysis further demonstrates the need of finding ways for improving the accuracy of the prediction. To this end we developed a new protocol based on molecular structure similarity search and employing widely used solubility prediction models as implemented in free or commercial software. Similarity search

Table 2. Performance of the Solubility Prediction for All the Software on the Five Data Sets, According to Four Criteria: r^2, RMSE, Ratios of Correct Prediction P1 and P2

	Huuskonen				PhysProp				solubility challenge								SolDiv150			
									training				test							
models	r^2	P1	P2	RMSE	r^2	P1	P2	RMSE	r^2	P1	P2	RMSE	r^2	P1	P2	RMSE	r^2	P1	P2	RMSE
PP-solubility	0.86	0.49	0.79	0.84	0.40	0.38	0.64	1.34	0.40	0.32	0.55	1.34	0.29	0.15	0.42	1.66	0.60	0.45	0.73	1.03
PP-ADMET solubility	0.83	0.47	0.77	0.90	0.57	0.36	0.61	1.37	0.49	0.27	0.59	1.14	0.28	0.35	0.46	1.49	0.58	0.39	0.62	1.36
ACD	0.92	0.67	0.91	0.61	0.73	0.54	0.81	0.91	0.25	0.52	0.70	1.13	0.53	0.16	0.60	1.14	0.79	0.51	0.83	0.82
MOE	0.91	0.56	0.87	0.67	0.53	0.43	0.72	1.13	0.38	0.42	0.73	0.95	0.44	0.27	0.58	1.30	0.62	0.50	0.79	1.00
QikProp	0.88	0.63	0.84	0.76	0.54	0.40	0.66	1.21	0.37	0.46	0.72	1.03	0.34	0.31	0.65	1.24	0.76	0.49	0.77	0.89
QikProp-CI	0.84	0.52	0.80	0.86	0.51	0.40	0.69	1.24	0.26	0.35	0.66	1.11	0.57	0.46	0.77	1.12	0.72	0.42	0.72	0.97
ADMET PREDICTOR	0.94	0.75	0.94	0.51	0.64	0.53	0.81	0.96	0.45	0.45	0.79	0.96	0.32	0.23	0.65	1.25	0.73	0.59	0.90	0.90
Volsurf +	0.85	0.45	0.77	0.88	0.55	0.33	0.61	1.28	0.52	0.39	0.73	0.88	0.44	0.31	0.62	1.09	0.65	0.39	0.62	1.21
FAF	0.82	0.39	0.71	0.97	0.39	0.38	0.67	1.18	0.46	0.58	0.77	0.87	0.18	0.38	0.62	1.19	0.56	0.43	0.72	1.04
VCC lab	0.79	0.39	0.68	1.08	0.68	0.52	0.79	0.94	0.25	0.38	0.70	0.97	0.39	0.31	0.69	1.18	0.74	0.49	0.73	1.05
ISIDA	0.93	0.67	0.90	0.59	0.52	0.36	0.63	1.24	0.39	0.49	0.73	0.85	0.35	0.54	0.69	1.09	0.29	0.33	0.64	1.24

dx.doi.org/10.1021/mp300234q | Mol. Pharmaceutics XXXX, XXX, XXX–XXX

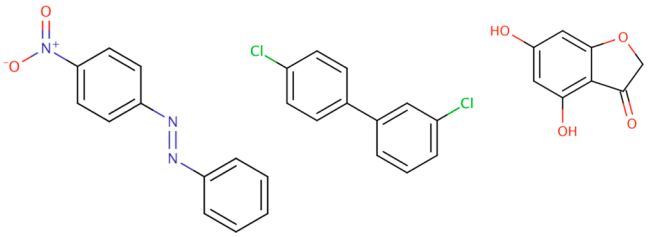


Figure 4. Three fragments found to be badly predicted by ISIDA.

of compounds has already been exploited for solubility prediction.[11,46] The common idea is to select the model which achieves the best solubility prediction for the most similar compounds of the targeted one. Our approach brings the following additional improvements: an algorithmic optimization through the implementation of a Random Forest scheme in order to find the best performing fingerprint for similarity search used for solubility prediction (see Figure 1S in the Supporting Information).

Keys To Select the Most Appropriate Model. In order to choose appropriate fingerprints allowing a satisfactory molecular similarity search, we investigated which fingerprints represented the best solubility. To address this problem, several models based on different molecular fingerprints were created using a nonparametric regression method for aqueous solubility prediction, a Random Forest algorithm (see in Experimental Section for details) shown previously to be suitable for solubility prediction.[47] The data set used for the training was the PhysProp data set with the largest number of compounds. Five fingerprints were compared: MACCS keys, ECFP and FCFP, both from length of 4 and 6.

The PhysProp data set was split into 7 folds of 500 random compounds, but the last, 8th fold contained the 457 remaining compounds. To avoid overfitting and provide a robust model, which can be applied to new data, an 8-fold cross validation was performed. This procedure divides the data set into 8 subsamples, learns a model using 7 subsamples and tests it on the remaining set. This iterative procedure is then repeated until each fold has served as test set. At the end of the process, the whole sample has been used as test and it is possible to compute the global model quality quantified by r^2. This procedure was repeated for each value of the parameters *ntree* and *mtry* (see Experimental Section), and those final values that

Table 3. Values of the r^2, RMSE and Ratio of Correct Prediction (P1 and P2) Obtained for Each Fingerprint

fingerprint	ECFP_6	ECFP_4	FCFP_6	FCFP_4	MACCS
r^2	0.55	0.56	0.58	0.59	0.68
P1	0.39	0.40	0.40	0.41	0.48
P2	0.66	0.67	0.70	0.71	0.77
RMSE	1.20	1.18	1.15	1.13	0.99

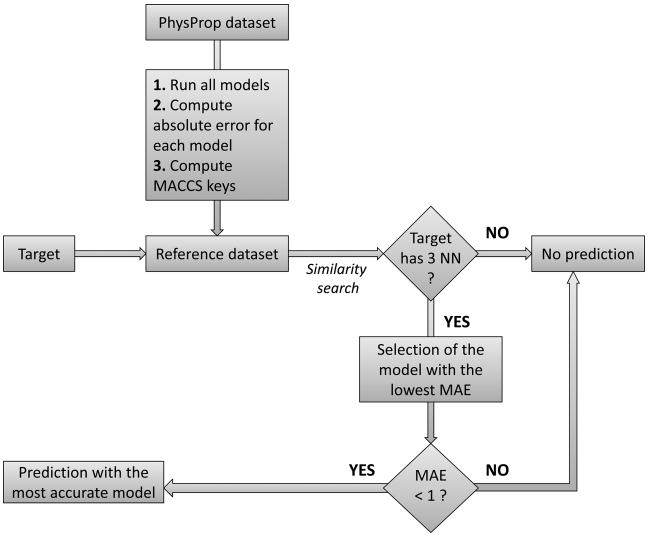


Figure 6. Schematic description of the multimodel protocol.

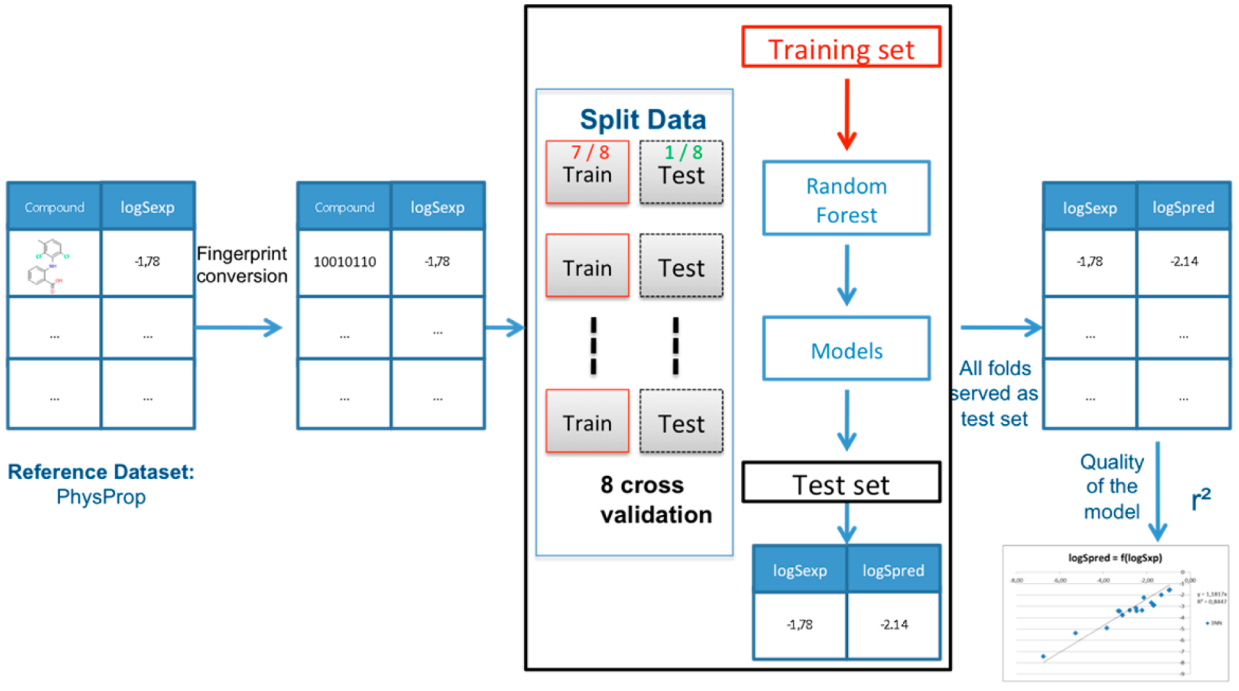


Figure 5. Schematic description of the Random Forest algorithm for the selection of the best fingerprint.

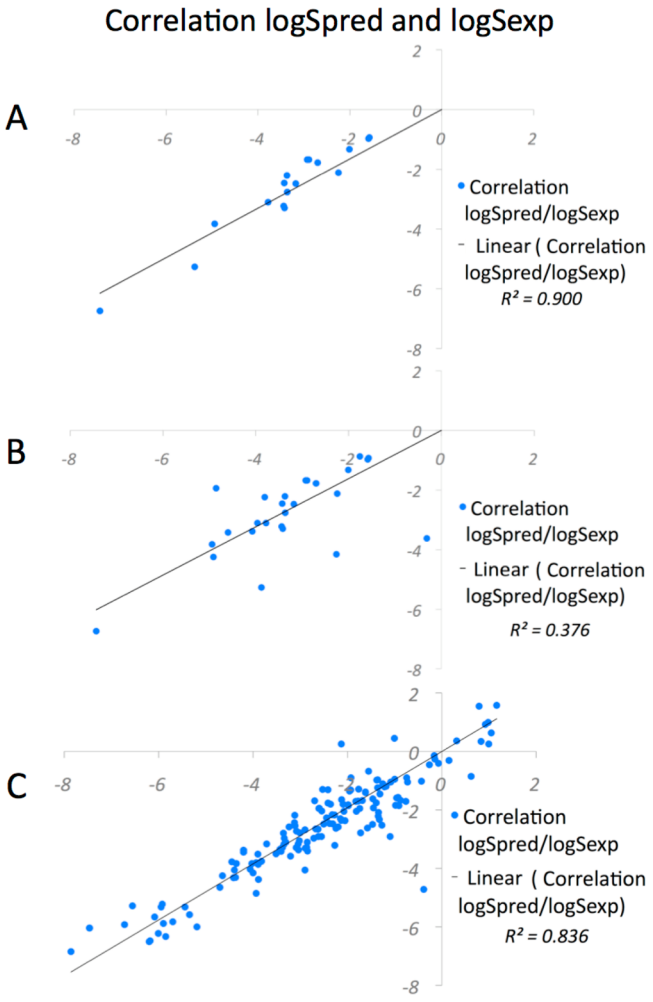


Figure 7. Correlation between the predicted and experimental logS. (A) For the 17 compounds of the Solubility Challenge test set satisfying the similarity cutoff of 0.7. (B) For the 26 compounds of the Solubility Challenge test set when using the similarity cutoff of 0.5. (C) For the 150 compounds of the SolDiv150 data set satisfying the similarity cutoff of 0.7.

Table 4. Performance of Models on 62 Drugs Extracted from the PhysProp Data Set, That Have 3 Similar Compounds in the Reference Set

models	P1	P2	RMSE	r^2
PP-solubility	0.59	0.87	0.77	0.53
PP-ADMET solubility	0.43	0.73	0.95	0.51
ACD	0.56	0.84	0.87	0.48
MOE	0.68	0.92	0.51	0.62
QikProp	0.67	0.84	0.86	0.45
QikProp-CI	0.56	0.75	0.88	0.30
ADMET Predictor	0.54	0.86	0.72	0.56
VolSurf +	0.49	0.81	0.88	0.13
FAF	0.46	0.86	0.85	0.39
VCC lab	0.43	0.78	0.75	0.51
ISIDA	0.40	0.76	0.89	0.23
multimodel	0.62	0.90	0.75	0.58

led to the maximal cross-validated r^2 were chosen. This protocol is illustrated in Figure 5.

Finally, 5 new models, one for each fingerprint, were trained using the same 8-fold cross validation techniques with the optimized parameters. The results of the r^2, RMSE and ratio of correct prediction for each model are presented in Table 3.

According to the results, the MACCS keys are best fingerprint descriptors among the five tested for aqueous solubility prediction, and thus we decided to choose MACCS keys for the selection of most similar compounds in our multimodel protocol.

Multimodel Protocol Scheme. For a target compound, our protocol selects the model providing the best solubility prediction for structurally similar compounds. We used a k = 3 nearest neighbors approach based on a structure similarity measure. This value was chosen because we obtained the best r^2 for k = 3 when varying k from 1 to 6 (tests performed on the Solubility Challenge test set as containing reliable experimental data; shown in Figure S2 in Supporting Information). The three most similar compounds are taken into account according to the Tanimoto similarity coefficient of 0.7 using the MACCS keys fingerprints. A schematic description of the algorithm is given in Figure 6 and Figure 1S in the Supporting Information. The entire protocol was implemented in Pipeline Pilot. The PhysProp data set was taken as a reference for the experimental solubility data. The protocol chooses the model achieving the minimal mean error MAE for the three reference compounds. If the best chosen model exhibits a mean error of more than 1 log, no prediction is proposed. We also explored the possibility to employ the protocol in case of absence of 3 compounds with similarity values of 0.7 by decreasing this barrier to 0.5. However, in such cases, the performance is not satisfactory as it is shown below. In fact, missing experimental solubility data for compounds similar to the target molecule indicates that it may not belong to the applicability domain and limits significantly the applicability of the method leading to worse prediction as already observed in ref 11.

External Validation of Multimodel Method. *Solubility Challenge Test.* Although the number of compounds in this data set is small (26 compounds), this data set was chosen because most of the models presented in the benchmarking section of our study performed poorly. Seventeen compounds satisfied the triple *cutoff* barrier of 0.7 similarity, which represent 65% of the initial data. The obtained r^2 of 0.90 suggests a good performance (illustrated in Figure 7A), as compared to the best individual model QikProp-CI applied to the same 17 compounds and achieving r^2 equal to 0.76.

Next, we tested the performance with the 26 compounds and a cutoff of 0.5 in terms of similarity in the cases where no sufficiently similar compounds were present in the reference experimental data set (Figure 7B). In this case we observed an important decrease of the performance. Indeed, the r^2 is equal to 0.38 and outliers appear on the graph. This is due to the fact that for some molecules the prediction is based on molecules structurally different from the compound of interest. This points out the importance of verifying if the compound of interest belongs to the applicability domain of the used method.

SolDiv150. In order to validate the multimodel protocol on a larger number of compounds, the 150 compounds of SolDiv150 were then tested. These 150 compounds were removed temporarily from the reference data set. The results are represented in Figure 7C. A significant improvement can be observed in comparison with the individual models on this data set (see Table 2). Indeed, the best performing model on the SolDiv150 set, ACD, achieved r^2 of 0.79. After applying the multimodel protocol we increased the r^2 to 0.84. ADMET Predictor showed a ratio of correct prediction P1 of 0.59 on

dx.doi.org/10.1021/mp300234q | Mol. Pharmaceutics XXXX, XXX, XXX–XXX

SolDiv150, while our protocol improved it to 0.63. The RMSE is also slightly decreased since ACD obtained 0.82 and our protocol 0.71. Overall, our protocol combining different models for solubility prediction gave more accurate results than the individual ones by enlarging the applicability domain. The most important value of such an approach is the possibility to determine which is the most appropriate model that can be used on specific chemotypes. In fact, the applicability domain of our method depends on the applicability domains of each individual QSAR model. One can speculate that our approach could also be employed to other available solubility models while implementing larger in-house reference data sets.

Drugs. In order to assess the prediction performance on real drugs, we repeated the evaluation analysis for extracted drugs from PhysProp; we found 62 molecules that have 3 similar compounds in the reference set. As can be seen from Table 4, all models perform unsatisfactorily on this drug data set, in terms of r^2, compared to the druglike compounds used. The best performing approaches are MOE and the multimodel protocol. Detailed information for solubility prediction of these two models can be found in Table 1S in the Supporting Information. However, MOE performs poorly on PhysProp, both Solubility Challenge data sets and SolDiv150. As we obtained the best results on the drugs' data set with MOE, it might be suggested drug molecules could be well represented in its training set, which is apparently not the situation for the other models. Regarding the multimodel protocol, for only 6 drugs the absolute error of logS > 1 log unit, and only for 1 drug the absolute error > 2 log units, which is still acceptable.

CONCLUSION

We have presented a new protocol that improves solubility prediction based on optimized structure similarity search, data sets and the combination of several packages. Our multimodel protocol allows selecting the most accurate model for a compound of interest among several possible models. We evaluated the performance of different fingerprints for aqueous solubility prediction based on molecular structure similarity through a Random Forest approach. Our analysis demonstrated that the multimodel protocol significantly improves solubility prediction on a large number of diverse compounds as compared to the ten individual models thoroughly assessed here. Yet, additional improvement is needed for real drugs. We note that missing experimental solubility data for compounds similar to the target compound/drug generally limit significantly the applicability of the method leading to worse prediction. Yet, these compounds are flagged and it is then possible to measure experimentally solubility for these molecules such as to maintain the high level of accuracy prediction seen for the other compounds.

ASSOCIATED CONTENT

Supporting Information

Figure 1S, visualization of the developed multimodel protocol; Figure 2S, performance of the multimodel protocol depending on the numbers of nearest neighbors (NN) as applied to the Solubility Challenge test set; Table 1S, experimental and predicted solubility data for 62 drugs extracted from PhysProp; SDF file containing the 150 molecules of SolDiv150.

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

AUTHOR INFORMATION

Corresponding Author

*P.V.: e-mail, philippe.vayer@fr.netgrs.com; tel, +33 238 238 003. M.A.M.: Université Paris Diderot, Sorbonne Paris Cité, Molécules Thérapeutiques *in silico*, Inserm UMR-S 973, 35 rue Helene Brion, 75013 Paris, France; e-mail, maria.miteva@univ-paris-diderot.fr; tel, +33157278392; fax, +33157278372.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Support from Servier, the French National Research Institute Inserm, and the University Paris Diderot is greatly appreciated. We thank Schrödinger for providing the evaluation version of the QikProp program and the Simulations Plus company for providing the evaluation version of the ADMET Predictor software.

REFERENCES

(1) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(2) Di, L.; Kerns, E. H. Biological assay challenges from compound solubility: strategies for bioassay optimization. *Drug Discovery Today* **2006**, *11*, 446−451.

(3) Alelyunas, Y. W.; Empfield, J. R.; McCarthy, D.; Spreen, R. C.; Bui, K.; Pelosi-Kilby, L.; Shen, C. Experimental solubility profiling of marketed CNS drugs, exploring solubility limit of CNS discovery candidate. *Bioorg. Med. Chem. Lett.* **2010**, *20*, 7312−7316.

(4) Ishikawa, M.; Hashimoto, Y. Improvement in Aqueous Solubility in Small Molecule Drug Discovery Programs by Disruption of Molecular Planarity and Symmetry. *J. Med. Chem.* **2011**, *54*, 1539−1554.

(5) Wang, J.; Hou, T. Recent Advances on Aqueous Solubility Prediction. *Comb. Chem. High Throughput Screening* **2011**, *14*, 328−338.

(6) van de Waterbeemd, H. Improving Compound Quality through in vitro and in silico Physicochemical Profiling. *Chem. Biodiversity* **2009**, *6*, 1760−1766.

(7) Wang, J.; Krudy, G.; Hou, T.; Zhang, W.; Holland, G.; Xu, X. Development of Reliable Aqueous Solubility Models and Their Application in Druglike Analysis. *J. Chem. Inf. Model.* **2007**, *47*, 1395−1404.

(8) Hopfinger, A. J.; Esposito, E. X.; Llinas, A.; Glen, R. C.; Goodman, J. M. Findings of the Challenge To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2008**, *49*, 1−5.

(9) Llinas, A.; Glen, R. C.; Goodman, J. M. Solubility Challenge: Can You Predict Solubilities of 32 Molecules Using a Database of 100 Reliable Measurements? *J. Chem. Inf. Model.* **2008**, *48*, 1289−1303.

(10) Hewitt, M.; Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Roberts, D. W.; Dearden, J. C. In Silico Prediction of Aqueous Solubility: The Solubility Challenge. *J. Chem. Inf. Model.* **2009**, *49*, 2572−2587.

(11) Kuhne, R.; Ebert, R.-U.; Schormann, G. Model Selection Based on Structural Similarity: Method Description and Application to Water Solubility Prediction. *J. Chem. Inf. Model.* **2006**, *46*, 636−641.

(12) Delaney, J. S. Predicting aqueous solubility from structure. *Drug Discovery Today* **2005**, *10*, 289−295.

(13) Ran, Y.; Yalkowsky, S. H. Prediction of Drug Solubility by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 354−357.

(14) Yalkowsky, S. H.; Valvani, S. C. Solubility and partitioning I: Solubility of nonelectrolytes in water. *J. Pharm. Sci.* **1980**, *69*, 912−922.

(15) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234−252.
(16) Sanghvi, T.; Jain, N.; Yang, G.; Yalkowsky, S. H. Estimation of Aqueous Solubility By The General Solubility Equation (GSE) The Easy Way. *QSAR Comb. Sci.* **2003**, *22*, 258−262.
(17) Wassvik, C. M.; Holmen, A. G.; Draheim, R.; Artursson, P.; Bergstrom, C. A. Molecular characteristics for solid-state limited solubility. *J. Med. Chem.* **2008**, *51*, 3035−3039.
(18) Ali, J.; Camilleri, P.; Brown, M. B.; Hutt, A. J.; Kirton, S. B. Revisiting the General Solubility Equation: In Silico Prediction of Aqueous Solubility Incorporating the Effect of Topographical Polar Surface Area. *J. Chem. Inf. Model.* **2012**, *52*, 420−428.
(19) McElroy, N. R.; Jurs, P. C. Prediction of Aqueous Solubility of Heteroatom-Containing Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1237−1247.
(20) McFarland, J. W.; Avdeef, A.; Berger, C. M.; Raevsky, O. A. Estimating the Water Solubilities of Crystalline Compounds from Their Chemical Structures Alone. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1355−1359.
(21) Butina, D.; Gola, J. M. R. Modeling Aqueous Solubility. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 837−841.
(22) Hou, T.; Wang, J. Structure - ADME relationship: still a long way to go? *Expert Opin. Drug Metab. Toxicol.* **2008**, *4*, 759−770.
(23) Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **2003**, *22*, 69−77.
(24) Jorgensen, W. L.; Duffy, E. M. Prediction of drug solubility from structure. *Adv. Drug Delivery Rev.* **2002**, *54*, 355−366.
(25) Lagorce, D.; Maupetit, J.; Baell, J.; Sperandio, O.; Tuffery, P.; Miteva, M. A.; Galona, H.; Villoutreix, B. O. The FAF-Drugs2 server: a multistep engine to prepare electronic chemical compound collections. *Bioinformatics* **2011**, *27*, 2018−2020.
(26) Cheng, A.; Merz, K. M. Prediction of Aqueous Solubility of a Diverse Set of Compounds Using Quantitative Structure Property Relationships. *J. Med. Chem.* **2003**, *46*, 3572−3580.
(27) Stuart, M.; Box, K. Chasing Equilibrium: Measuring the Intrinsic Solubility of Weak Acids and Bases. *Anal. Chem.* **2005**, *77*, 983−990.
(28) PhysProp. http://www.syrres.com/, 2012.
(29) Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous Solubility Prediction of Drugs Based on Molecular Topology and Neural Network Modeling. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 450−456.
(30) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.
(31) Hou, T. ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 266−275.
(32) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5−32.
(33) R Development Core Team. R Foundation for Statistical Computing. *R: A language and environment for statistical computing*. http://www.R-project.org, 2005.
(34) McGregor, M. J.; Pallai, P. V. Clustering of Large Databases of Compounds: Using the MDL Keys as Structural Descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.
(35) Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1000−1005.
(36) Klopman, G.; Wang, S.; Balthasar, D. M. Estimation of aqueous solubility of organic molecules by the group contribution approach. Application to the study of biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 474−482.
(37) Klopman, G.; Zhu, H. Estimation of the Aqueous Solubility of Organic Molecules by the Group Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 439−445.
(38) MOE. Chemical Computing Group, Inc., 2010.
(39) ACD/Labs. http://www.acdlabs.com/products/pc_admet/physchem/physchemsuite/, 2012.
(40) Schrodinger. http://www.schrodinger.com/products/14/17/, 2009.
(41) Simulation-Plus. http://www.simulations-plus.com/Products.aspx?grpID=1&cID=11&pID=13, 2012.
(42) Cruciani, G.; Pastor, M.; Guba, W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11* (Suppl. 2), S29−S39.
(43) Volsurf. http://www.moldiscovery.com, 2011.
(44) Tetko, I. V.; Tanchuk, V. Y. Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1136−1145.
(45) Infochimie, L. *ISIDA*. www.infochim.u-strasbg.fr, 2012.
(46) Zhang, H.; Ando, H. Y.; Chen, L.; Lee, P. H. On-the-fly selection of a training set for aqueous solubility prediction. *Mol. Pharmaceutics* **2007**, *4*, 489−497.
(47) 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*, 150−158.