Individual Hydrogen-Bond Strength QSPR Modelling with ISIDA Local Descriptors: a Step Towards Polyfunctional Molecules

ARTICLE · JUNE 2014

DOI: 10.1002/minf.201400032

Fiorella Ruggiu

University of Strasbourg

5 PUBLICATIONS **44** CITATIONS

SEE PROFILE

Gilles Marcou

University of Strasbourg

55 PUBLICATIONS **770** CITATIONS

SEE PROFILE

Dragos Horvath

French National Centre for Scientific Resea...

95 PUBLICATIONS **1,464** CITATIONS

SEE PROFILE

Jerome Graton

University of Nantes

52 PUBLICATIONS **774** CITATIONS

SEE PROFILE

DOI: 10.1002/minf.201400032

Individual Hydrogen-Bond Strength QSPR Modelling with ISIDA Local Descriptors: a Step Towards Polyfunctional Molecules

Fiorella Ruggiu,[a] Vitaly Solovev,[b] Gilles Marcou,[a] Dragos Horvath,[a] Jérôme Graton,[c] Jean-Yves Le Questel,[c] and Alexandre Varnek*[a]

Abstract: Here, we introduce new ISIDA fragment descriptors able to describe "local" properties related to selected atoms or molecular fragments. These descriptors have been applied for QSPR modelling of the H-bond basicity scale pK_{BHX}, measured by the 1:1 complexation constant of a series of organic acceptors (H-bond bases) with 4-fluorophenol as the reference H-bond donor in CCl_4 at 298 K. Unlike previous QSPR studies of H-bond complexation, the models based on these new descriptors are able to predict the H-bond basicity of different acceptor centres on the same polyfunctional molecule. QSPR models were obtained using support vector machine and ensemble multiple linear regression methods on a set of 537 organic compounds including 5 bifunctional molecules. They were validated with cross-validation procedures and with two external test sets. The best model displays good predictive performance on a large test set of 451 mono- and bifunctional molecules: a root-mean squared error *RMSE* = 0.26 and a determination coefficient $R^2 = 0.91$. It is implemented on our website (http://infochim.u-strasbg.fr/webserv/VSEngine.html) together with the estimation of its applicability domain and an automatic detection of potential H-bond acceptors.

Keywords: Equilibrium constants of hydrogen bonding · Hydrogen-bond acidity and basicity · Fragment descriptors · H-bond · ISIDA · pK_{BHX} · QSPR

1 Introduction

The hydrogen bond (H-bond) is one of the fundamental interactions between molecules and is of paramount importance for many properties, as well as for processes of living and abiotic nature. Many hydrogen-bonding effects are known such as density differences between ice and liquid water, joining cellulose microfibrils in wood, shaping DNA into genes and polypeptide chains into wool, hair, muscles or enzymes.[1]

The term hydrogen bond has been used for over a century but its precise definition and the nature of the interaction has been and is still stirring many debates. The IUPAC proposes the following short definition in a recent report:[2] "*The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment X–H in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation.*". Criteria for evidence of bond formation include the linearity and directionality of the interaction, spectroscopic evidence with a red-shift of the X–H vibrational frequency in infrared (IR) spectroscopy or the deshielding of the H in X–H in nuclear magnetic resonance spectroscopy and the thermodynamic characterization by the Gibbs free energy ΔG°.

In the case of intermolecular interactions, a hydrogen bond is formed between the molecule containing the X–H, referred as the H-bond donor (HBD), and the molecule containing the atom with which the X–H forms a bond, referred as the H-bond acceptor (HBA). These terms come from the electronic aspect of the hydrogen bond where the H-bond donor/acceptor is considered analogous to a Lewis acid/base. Hence, the term basicity is used to express the H-bond acceptors' strength which can be measured by thermodynamic quantities such as the equilibrium constant, the free energy, the enthalpy and the entropy of

[a] *F. Ruggiu, G. Marcou, D. Horvath, A. Varnek*
Laboratoire de Chémoinformatique, UMR 7140 CNRS, Université de Strasbourg
1, rue Blaise Pascal, 67000 Strasbourg, France
phone: +33368851560
**e-mail: varnek@unistra.fr*

[b] *V. Solovev*
Institute of Physical Chemistry and Electrochemistry, Russian Academy of Sciences
Leninskiy prospect, 31a, 119991, Moscow, Russian Federation

[c] *J. Graton, J.-Y. Le Questel*
Université de Nantes, UMR CNRS 6230, Chimie Et Interdisciplinarité: Synthèse, Analyse, Modélisation (CEISAM), UFR Sciences & Techniques
2, rue de la Houssinière, BP 92208, 44322 NANTES Cedex 3, France

Supporting Information for this article is available on the WWW under http://dx.doi.org/10.1002/minf.201400032.

the association reaction (see Equation 1) where HBA···HBD is the 1:1 formed complex.

$$\mathrm{HBA} + \mathrm{HBD} \leftrightarrow \mathrm{HBA} \cdots \mathrm{HBD} \qquad (1)$$

The H-bond is also essential in protein-drug interactions[3] and it seems that the evaluation of the strength of such an interaction is not intuitive for medicinal chemist.[4,5] Hence, a quantitative assessment of H-bond strength has for a long time been important for the chemical community. H-bond basicity scales, pK_{HB} and log K_{β},[6] have been constructed, based on the equilibrium constants of 1:1 complexation of a reference HBD, 4-fluorophenol and 4-nitrophenol, respectively, with different acceptors in CCl_4. The pK_{HB} scale was later extended to the pK_{BHX} scale by Laurence et al.[4] using Fourier transform IR (FTIR) spectroscopy. In Equation 2, the pK_{BHX} value is defined as the logarithm of the equilibrium constant, measured at 298 K, with 4-fluorophenol as the reference HBD. A strong advantage of FTIR measurements is found for the study of polyfunctional compounds, for which the significant multiple H-bond sites are observable and a measured pK_{BHX} value can be attributed to each site. To our knowledge, with around 1200 values, the pK_{BHX} database[4] constitutes the largest collection of data on H-bond basicity. Moreover, the diversity of H-bond acceptor functional groups encountered in the pK_{BHX} database enables the building of QSPR models with an expected large applicability domain.

$$pK_{BHX} = \log K = \log([\mathrm{HBA} \cdots \mathrm{HBD}]/[\mathrm{HBA}][\mathrm{HBD}]) \qquad (2)$$

Earlier, the modelling of thermodynamic parameters of the 1:1 H-bond complexes has already been attempted through various approaches such as quantum chemical methods,[7–12] linear free-energy relationships (LFERs),[13–17] empirical correlations[18–24] and quantitative structure-property relationships (QSPR) using results of quantum chemical calculations as descriptors.[7–11, 25–28] The LFERs models by Raevsky[15–17] and Abraham[14] consider the free energy of H-bond complexation as a product of the acceptor and donor parameters. To our knowledge, they were not properly validated except once on a small test set including 6 reactions.[15] QSPR models by Henneman et al.[26] for pK_{HB} based on AM1-calculated descriptors were obtained considering a limited set of 42 aromatic N-heterocycles and validated on a small set of 17 compounds, resulting in a mean absolute error of 0.17 log K units. The models by Besseau et al.[7] based on density functional theory approach were trained on 59 monofunctional nitrogen bases and validated on an external test set of 142 compounds with a root mean-squared error (*RMSE*) of 0.13 (calculated from the data reported in[7]). Klamt et al.[29] used the COSMO-RS approach to assess experimental H-bond enthalpies and free energies of about 300 H-Bond complexes from the pK_{BHX} database with an accuracy of ±2 kJ mol^{-1} (±0.35 log K units). In another recent study, Kerdawy et al.[12] performed a series of density functional/basis set combinations and second-order Møller–Plesset calculations on the complexes of 58 simple HBAs, mainly pyridine nitrogen and carbonyl oxygen sites, from the pK_{BHX} database, with methanol as HBD. A partial optimisation of the H-bond complex leads to reasonable correlations between pK_{BHX} and the calculated interaction energies, but no validation on an external test set was reported. Green et al.[28] found reasonable linear correlations (R^2_{corr}=0.91–0.97) between pK_{BHX} measured for 41 HBAs with quantum chemical topology descriptors calculated for the complexes of these compounds with 5 different HBDs (water, methanol, 4-fluorophenol, serine and methylamine). The correlation equation for methanol was successfully used to assess pK_{BHX} values for 11 bifunctional HBAs.

In our previous publication,[30] the ISIDA fragment descriptors were used to model free energy (ΔG°) and enthalpy (ΔH°) of the 1:1 complexes between organic acids and bases linked by one H-bond. In these complexes, the acids were substituted phenols, whereas the bases were represented by the large variety of chemical classes: phenols, alcohols, ethers, ketones, amides, heterocyclic compounds, phosphoryl and sulfonyl compounds. The ensemble Multiple Linear Regression (MLR) model built on a training set of 292 complexes was validated on a test set of 66 complexes. A reasonable correlation between predicted and experimental values was observed:

$$\Delta G_{pred} = 0.10 + 1.00\Delta G_{exp} \; (R^2{}_{corr} = 0.92, \mathit{RMSE} = 1.64\ \mathrm{kJ\,mol^{-1}}$$

i.e., 0.29 log K units).

Nowadays, mostly polyfunctional molecules (see Figure 1 as an example) are used to design new nanomaterials based on a network of H-bonds or in the drug design area. Attempts to assess simultaneously pK_{BHX} values of their different sites are still scarce. Quantum chemical models (e.g.

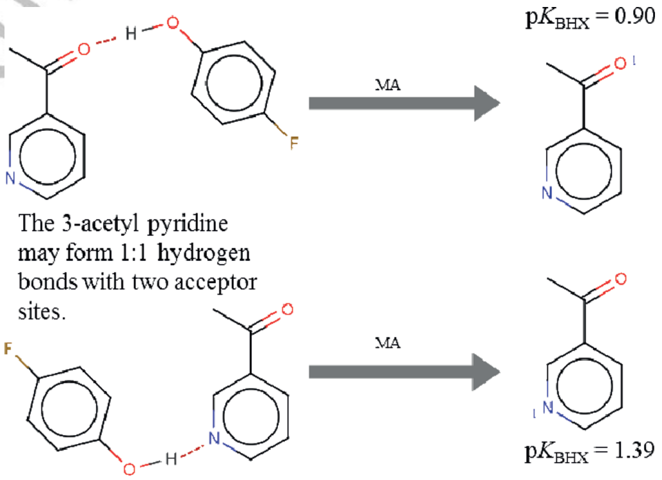

pK_BHX = 0.90
O
H
O
MA
O
1
F
N
N
The 3-acetyl pyridine may form 1:1 hydrogen bonds with two acceptor sites.
O
F
MA
O
N
H
O
1
N
pK_BHX = 1.39

Figure 1. Example of marked atom (MA) assignment on 3-acetyl pyridine. The marking is indicated with a 1 next to the acceptor's centre.

Literature[8–13]) can potentially be used for this purpose, in support of experimental measurements, to characterise the individual H-bond basicity of acceptor sites encountered in progesterone,[31] cotinine,[32] lobeline,[33] myosmines,[34] codeine and galanthamine.[35] Performance of quantum chemical topology descriptors to treat polyfunctional molecules have also been demonstrated by Green et al.[28] At the same time, heavy and time-consuming quantum mechanics-based approaches could hardly be recommended for virtual screening of large databases frequently used in computer-aided drug- or material design. The need for fast and reliable QSPR approaches able to assess HBD or acceptor ability of different binding sites of polyfunctional molecules is thus obvious. QSPR modelling is closely related to the question of models' applicability domain. Up to now, the best performing models were built on small data sets containing very specific acceptor chemotypes which certainly limited their application to similar compounds.

In extension of our previous work,[30] we here report new molecular descriptors specifically developed to model hydrogen-bonding parameters of polyfunctional molecules. These new descriptors have been used to build QSPR models for pK_{BHX} on a large structurally diverse dataset containing 537 mono- and bi-functional compounds. The models were validated on two external test sets containing 451 and 36 HBAs and were implemented on the web for the end users.

2 Computational Procedure

The general procedure followed in this work is summarized in the workflow shown in Figure 2. First the dataset is extracted from the pK_{BHX} database and processed in order to build different QSPR models. The resulting models were validated on two external test sets.

2.1 Data Preparation

Molecules from the pK_{BHX} database[4] were first filtered by removing all predicted values and by removing salts. An entry containing iron was also removed. They were then standardised by using an implicit representation of hydrogen, choosing a standard representation for groups such as nitro or imidazole, and generate major tautomer using ChemAxon's Calculator plugin.[36] The acceptors' sites had been identified experimentally during the measurement: they are indicated in the database and were thus easy to mark. If a compound contained several equivalent acceptor sites, only one of them was kept. The EdiSDF software[37] has been used to mark HBA centres.

The training set contains organic acceptors including different types of hydrogen-bonding atoms which have been categorised into 11 families:

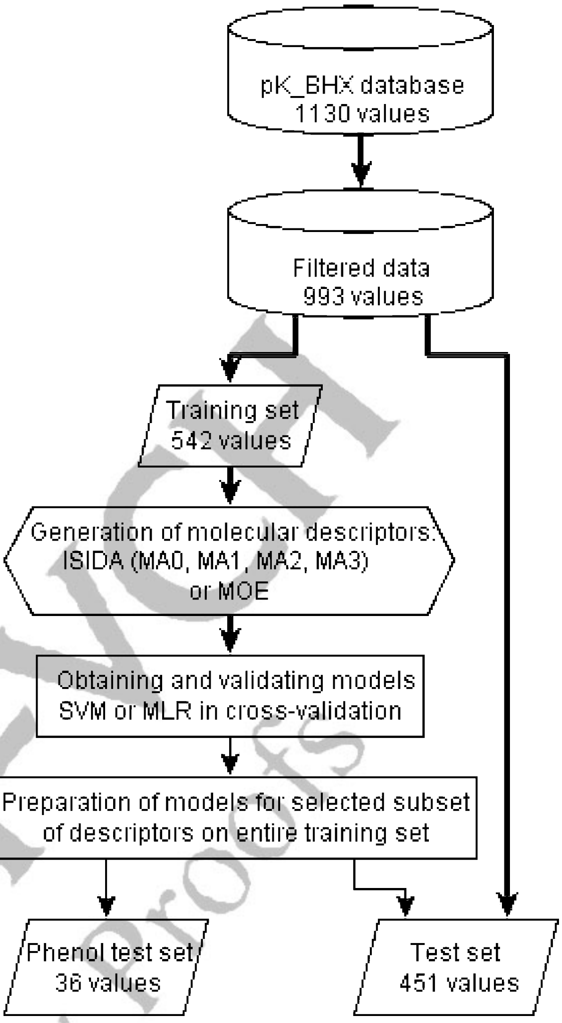

pK_BHX database
1130 values
Filtered data
993 values
Training set
542 values
Generation of molecular descriptors:
ISIDA (MA0, MA1, MA2, MA3)
or MOE
Obtaining and validating models
SVM or MLR in cross-validation
Preparation of models for selected subset
of descriptors on entire training set
Phenol test set
36 values
Test set
451 values

Figure 2. Workflow for the QSPR modelling of the pK_{BHX} database.

- the carbonyl oxygen (esters, carbonates, lactones, aldehydes, ketones, amides, lactams, carbamates, ureas),
- the ether oxygen (ethers, alcohols),
- the oxygens of nitro group (nitroalkenes, nitroaromatics, nitramides),
- the oxygen of sulfinyl group (sulfites, sulfoxides),
- the oxygen of phosphoryl group (phosphoramides, phosphine oxides, phosphonates, phosphates),
- the amine nitrogen (anilines, amines),
- the imine nitrogen (amidines, pyrrolines, imines),
- the aromatic nitrogen (pyridines, azoles),
- the nitrile nitrogen (nitriles),
- the sulfur of sulfide (sulfides),
- the sulfur of thiocarbonyl group (thioamides, thioureas, thiocarbonates, thioketones, isothiocyanates).

For five bifunctional compounds (3-acetylpyridine, morpholine, N-methylmorpholine, thiomorpholine, and thiazolidine), the pK_{BHX} values were specified for two different acceptor sites. For the molecules in the training set, pK_{BHX}

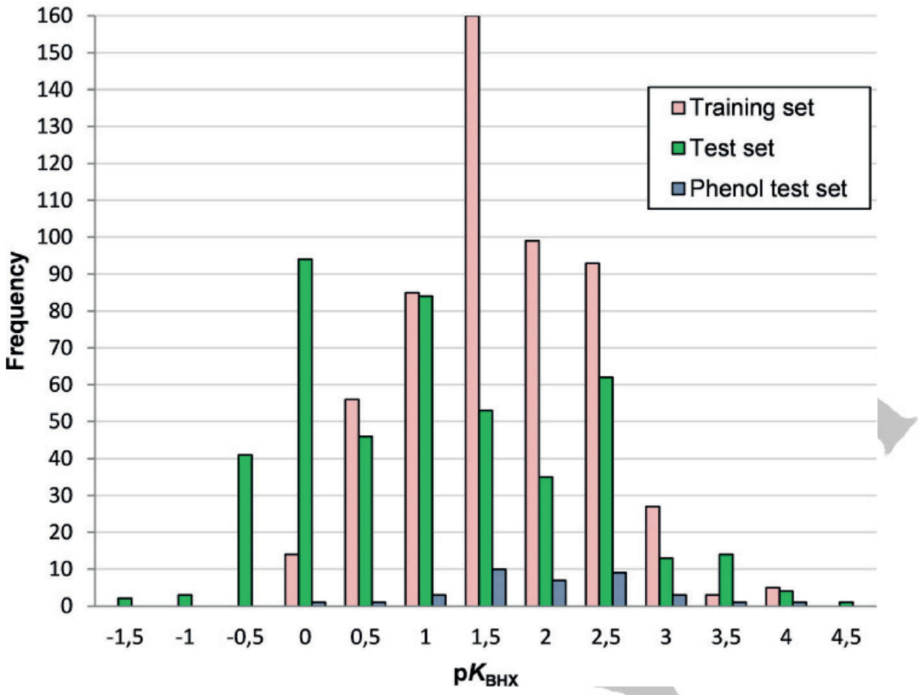


Figure 3. Experimental pK_{BHX} values distribution for the training set (red hatched), the test set (green) and the phenol set (blue dotted).

varies from −0.37 to 3.66 (Figure 3), whereas the size varies from 1 to 24 heavy atoms. Only direct experimental pK_{BHX} values were selected for the training set, using the exact same protocol for measurement. Thus, the training data are expected to have low internal errors, thereby ensuring the quality of data before the actual modelling. Random experimental error was evaluated by the experimentalists to be 0.04 log units.[7]

The training set is thus composed of a subset of 537 compounds (532 mono- and 5 bi-functional, in total 542 pK_{BHX} values) corresponding to direct and non-approximated experimental pK_{BHX} values. The remaining compounds, for which pK_{BHX} values were considered as less reliable, constituted an external test set (*test set*) of 451 values. The pK_{BHX} values in the latter were either corrected (when several equivalent acceptors' atoms are present) or approximated (for low soluble compounds, extreme values or values estimated from a measure of the IR shift of the donor's OH bond). Some compounds correspond to acceptor types not present in the training set (e.g. aromatic or alkene fragments, sulfates or halogens). These problematic molecules provide a good challenge to the models applicability domain. This set also includes 47 polyfunctional molecules with 2 non-equivalent acceptor sites.

One other supplementary external test set of 36 acceptors with phenol[38,39] (*phenol set*) (see Supporting information Section 3 Table SI4) was collected from the literature.

The labelling of the acceptor sites in the training and test sets have been performed manually according to database annotations. However, an automatic detection of acceptor sites has been implemented in the program to screen virtual libraries with developed models.

2.2 Descriptors

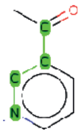
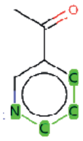
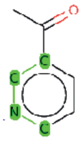
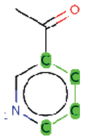
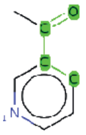
Two types of descriptors were used in this study: ISIDA fragment descriptors[30,40–42] and classical molecular descriptors generated by the Molecular Operating Environment (MOE) 2011 program.[43]

ISIDA fragment descriptors[30,40–42] represent subgraphs of a molecular graph. Each unique subgraph is considered as a descriptor, whereas its occurrence is used as the descriptor's value. The atoms of the molecular graph may be represented as their elements symbols, also called atom symbols, or forcefield typing according to the consistent valence forcefield[44] as other properties including pharmacophoric flagging and partial charge-based bins.41]

Once, the molecular graph has been represented with the desired property, it is then segmented using a particular fragmentation scheme: sequences and atom-centred fragments of varying length. The minimum length of fragments varied from 2 to 4 and the maximum length from 2 to 15. By default, the algorithm searches for the shortest possible path between two atoms but the all path exploration option has also been tried. The atom pair option was also conversely used and consists of representing the extremities of the fragment and the length of the path between them.

Our working hypothesis is that hydrogen-bonding acceptor strength is chiefly influenced by the accepting atom,

Table 1. Examples of sequence descriptors in the different classes of different sequence paths of length 4 in 3-acetyl pyridine with the N as marked atom. If the path is not represented in the description, the field is left empty.

					
MA0 – No marked atom, all fragments	N*C*C–C	N*C*C*C	C*N*C*C	C*C*C*C	O=C–C*C
MA1 – only fragments beginning with the marked atom	N*C*C–C	N*C*C*C			
MA2 – only fragments containing the marked atom	N*C*C–C	N*C*C*C	C*N*C*C		
MA3 – all fragments with a special flag on the marked atom	N&MA&*C*C–C	N&MA&*C*C*C	C*N&MA&*C*C	C*C*C*C	O=C–C*C

the heavy donor atom and the nature of their environment. For the dataset used in this work, the structure of the HBA is the only changing factor which influences the variation in pK_{BHX}. Therefore, we prepared special types of ISIDA fragment descriptors containing marked atoms (MA) which explicitly indicate the acceptor's atom position (Figure 1). In such a way, information about both acceptor centre and its environment is encoded.

Different marked atom strategies were considered (Table 1):

- No marked atom – all fragments are generated (MA0).
- Sequences start with the marked atom or the central atom of atom-centred fragments is the marked atom (MA1).
- Only fragments containing the marked atom are generated (MA2).
- A flag (&MA&) is added to the symbol of the marked atom and all fragments are generated (MA3).

A total of 1260 descriptor families were generated for each of the 4 tested marking strategies using either the ISIDA Fragmentor[45] or ISIDA/QSPR[46] programs. Each family contains from 10 to 10000 fragment descriptors. The labelling of the acceptor sites in the training and test sets have been performed manually according to database annotations. Notice that MA1 and MA2 descriptors were already suggested in our previous work.30] It should also to be noted that descriptors centred on selected atoms (different from those suggested in this work) were earlier applied in QSAR modelling of the dissociation constant (p*K*a).47–49]

2.3 MOE Descriptors

MOE descriptors were considered for the purpose of comparison. They represent a collection of 181 2D molecular descriptors, computed with the MOE 2011 program.[43] These descriptors describe physical properties, van der Waals surface area (the subdivided surface areas), the atom and bond counts (subdivided according to various criteria), the Kier and Hall chi connectivity and kappa molecular shape indices, the distance and adjacency matrices (Balaban's connectivity topological index, Wiener path number, Wiener polarity number), the pharmacophore atom types, and partial charges. All the hydrogen atoms were explicitly represented in the structures for the partial charge calculations.

2.4 Building and Validation of Models

Models were built and validated using support vector machines (SVM) with the LibSVM package[50] and MLR with the ISIDA/QSPR program.[46] Validation of models was carried out using cross-validation procedures (CV).[51] The determination coefficient (R^2), the correlation coefficient (R^2_{corr}) and the root mean squared error were used to evaluate the model ability to reproduce quantitatively the experimental data for training ($Y = Y_{calc}$) and test ($Y = Y_{pred}$) sets:

$$R^2 = 1 - \frac{\sum (Y_{exp} - Y)^2}{\sum (Y_{exp} - < Y_{exp} >)^2} \tag{1}$$

$$R^2_{corr} = \frac{\sum \{(Y_{exp} - < Y_{exp} >)(Y - < Y >)\}}{\sqrt{\sum (Y_{exp} - < Y_{exp} >)^2 \sum (Y - < Y >)^2}} \tag{2}$$

$$RMSE = \sqrt{\frac{\sum (Y - Y_{exp})^2}{n}} \tag{3}$$

where Y_{calc}, Y_{pred} and Y_{exp} are fitted, predicted and experimental values (here, $Y \equiv pK_{BHX}$) and summations are performed on all instances in the test set.

Ensemble modelling[52] implies the generation of many QSPR models, the selection of the most relevant ones and followed by their joint application to test compounds. For each compound from the test set, the program applies a *consensus model (CM)*, i.e., computes the property as an average of estimated values obtained with an ensemble of the models selected at the training stage. The standard deviation associated with this averaging can be used as a trustworthiness criterion: reliable predictions correspond to small standard deviations thus demonstrating that most of the models converge toward one same value.[53,54]

Any individual model entering the consensus model should be significantly better than *y*-scrambled models.[55] Therefore, for each of the employed machine learning techniques (SVM and MLR, see below) models' performances on 20 *y*-scrambling experiments were used to fit a normal distribution and a model was accepted only if its cross-validated performance was better than the 95th percentile of this distribution.

The SVM calculations were performed with epsilon support vector regression and a linear kernel using the LibSVM package.[50] Epsilon was optimised by fully exploring 10 values between 0.05 and 0.19 for the MA1 strategy. It was then set at 0.09 for the remaining of the study. The cost parameter was scanned (on a log scale) by 28 values ranging from 0.1 to 100. Each model was validated using a 5-fold cross-validation (5CV) with the inbuilt procedure of LibSVM with random splitting into learning/left-out subsets, which was reiterated 5 times in order to obtain robust average CV statistics. Models with 5CV $RMSE \leq 0.29$ pK_{BHX} units entered the CM. The *p*-value of the worse selected model compared to the y-scrambling performances was less than 0.001, thus far better than the minimum requirement to enter the CM. Only one SVM model (i.e. one cost parameter in this case) for each descriptor space was selected and then rebuilt on the entire training set to enter the CM.

Linear regression models were obtained with the ISIDA/QSPR software.[46] The individual MLR models were built by combining forward[56] and backward[57] stepwise variable selection techniques. In our calculations, many MLR models were generated combining different types of fragment descriptor and different variable selection algorithms. The number of generated individual models varied from 240 to 720 as a function of the descriptors used. Models with a leave-one-out CV determination coefficient $R^2 > 0.8$ were accepted for CM. All selected individual models performed significantly better than the scrambled models. Models were rebuilt on the entire training set for the CM.

Both SVM and MLR consensus models built on the entire training set were validated on two external tests sets (see Figure 2).

2.5 Applicability Domain (AD)

Generally, the AD[40] of the model defines an area of chemical space (basically, the one being densely covered by training set examples) where the model is presumably accurate. Two types of AD definitions were used simultaneously in this study: (i) *Fragment control* which consists in discarding predictions of test compounds containing fragments not occurring in the training set; (ii) *Bounding box* which considers AD as a multidimensional descriptor space confined by minimal and maximal values of occurrences of the descriptors involved in an individual model.

The applicability of a consensus model relies on the fraction of applicable individual models (i.e. the models for which AD does not discard the given molecule). If this number is lower than a threshold, the overall CM prediction is ignored.[58] In the ISIDA/QSPR software, by default, the threshold is 15% of the total number of models in the CM. For SVM, by default, the threshold is fixed at 50%.

3 Results and Discussion

3.1 Benchmarking of the Different Marked Atom Strategies

Individual SVM and MLR models were built using the different MA strategies and a Student's *t*-test was applied to the 5CV results to compare them. Average 5CV RMSE of the best models for each strategy are summarised in Table 2 and in Supporting Information Table SI 5. A *t*-test indicated that MA3 and MA2 descriptors are not significantly different at a confidence interval of 95%, while they are significantly different to MA1 and MA0 descriptors. Best models involving MA3 descriptors perform well in 5CV, both in SVM ($RMSE = 0.27$) and in MLR CM ($RMSE = 0.25$). The individual SVM models and the MLR CMs confirm the relevance of marked atom especially compared to 2D MOE descriptors ($RMSE = 0.40$, see Table 2).

Table 2. Predictive performances of the models in 5-fold cross-validation involving the different marked atom strategies and MOE descriptors.

Descriptors	Best individual SVM models		MLR CM	
	5CV-*RMSE*	5CV-R^2	5CV-*RMSE*	5CV-R^2
ISIDA MA0	0.33	0.80	0.35	0.80
ISIDA MA1	0.31	0.82	0.28	0.86
ISIDA MA2	0.28	0.86	0.27	0.87
ISIDA MA3	0.27	0.87	0.25	0.88
2D MOE	0.40	0.71	0.41	0.70

The first proposed approach MA0 does not pinpoint the hydrogen-bonding site and generically describes the molecule, similarly to MOE terms. MA1 and MA2 describe the immediate surroundings of the acceptor site. The last approach, MA3, can be seen as a combination of the two different points of view mentioned previously. It encompasses the whole molecule but adds the information of the HBA so that the machine learning procedures can differentiate atoms of the same type participating or not in hydrogen bonding.

MOE descriptors perform badly, but, surprisingly, MA0 performs well in SVM. This may be due to the small size of the organic molecules and the fact that, in most cases, only one easily identifiable acceptor is found. Thus, the machine learning seems able to identify some substructures related to the surroundings of the acceptor. Eventually, since the MA3 strategy seemed to perform better in general, it was preferred for ensemble modelling in SVM and in MLR for the prediction of the external test sets.

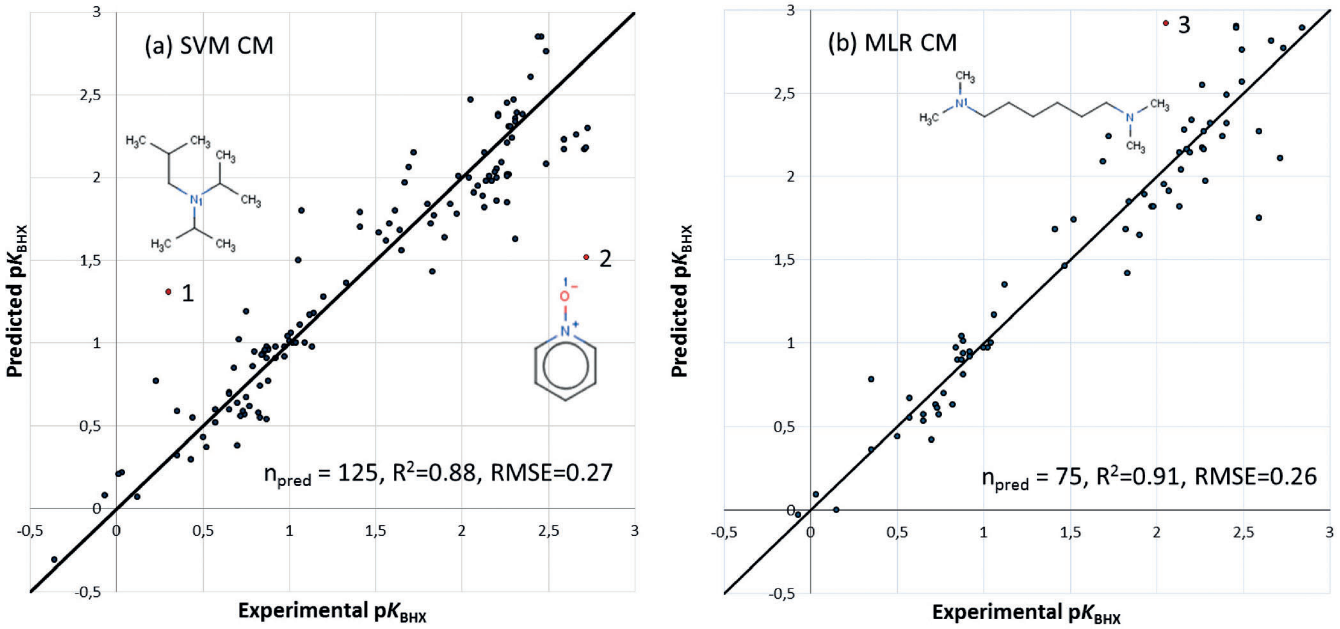


Figure 4. Predicted vs experimental pK_{BHX} values of the test set taking into account fragment control and bounding box as AD by (a) the CM SVM model where a minimum of 14 applicable models were required and (b) the CM MLR model. For the outliers 1–4, models accepted by AD greatly diverge thus showing low trustworthiness of predictions.

Notice that AD does not significantly change the models performance. Thus, MLR CM involving MA3 descriptors achieves a *RMSE* = 0.22 on 85 % of the data within AD.

3.2 Ensemble Modelling

In total, 27 models were selected for the SVM CM (see Supporting Information Table SI 6) and 306 for the MLR CM. The individual models were rebuilt on the entire training set using the sets of descriptors corresponding to the best individual models selected at the cross-validation step. Corresponding SVM and MLR CMs were applied to the external test set and the phenol set. Predicted pK_{BHX} values were assessed taking into account the fragment control and bounding box AD approaches, as well as the number of applied models.

Both SVM and MLR consensus models perform well on the test set: *RMSE* = 0.29 (SVM) and 0.26 (MLR), see Table 3 and Figure 4. These results are consistent with those obtained in cross-validation (see Table 2). The SVM model based on MOE descriptors performs poorly on the test set even if the SVM/ISIDA CM model's AD is accounted for: *RMSE* = 0.56 (Table 3).

The second external test set used was on 36 log *K* values with phenol[38,39] instead of 4-fluorophenol as HBD. According to literature,[59] such measurements should be highly correlated to pK_{BHX}. The MLR CM predictions are indeed correlated with R^2_{corr} = 0.86 with AD and SVM CM achieves R^2_{corr} = 0.92 with AD (see Supporting Information section 3 and Table SI 4–5). Thus, predicted pK_{BHX} and log *K* measured with phenol are highly correlated confirming previous observations by Raevsky et al.[16]

Table 3. Performance on the test set of the SVM and MLR consensus models (CM) based on ISIDA descriptors and the individual SVM model based on MOE descriptors using the SVM CM AD.

Method	Test set (with AD)			
	N_{mod} [a]	n_{pred} [b]	*RMSE*	R^2
SVM/ISIDA CM	14	125	0.27	0.88
	27	48	0.25	0.88
MLR/ISIDA CM	46	75	0.26	0.91
SVM/MOE	–	125	0.56	0.49

[a] N_{mod} is the minimum number of models required in CM. [b] n_{pred} is the number of acceptor sites accepted by AD.

In view of the better coverage of the phenol test set and its performance of R^2 of 0.91 on the test set, the MLR approach is considered marginally better than the SVM approach. It is interesting to note that in the associated linear correlation logK_{pred} (*p*-FC_6H_4OH) = −0.33 + 1.33 logK_{exp} (C_6H_5OH), the slope is superior to 1, thus confirming that the acidity of 4-fluorophenol is larger than that of phenol. Moreover, this linear correlation allows to predict from the phenol H-bond acidity (pK_{AHY} = 2.06, defined as the H-bond donor ability towards N-methylpyrrolidinone)[60] a value of 2.41 for 4-fluorophenol in excellent agreement with the experimental data (pK_{AHY} = 2.38).

In order to figure out how the CM performance depends on the number of applicable individual models, a series of SVM CM calculations have been performed on the test set. Figure 5 demonstrates variations of *RMSE* and the number

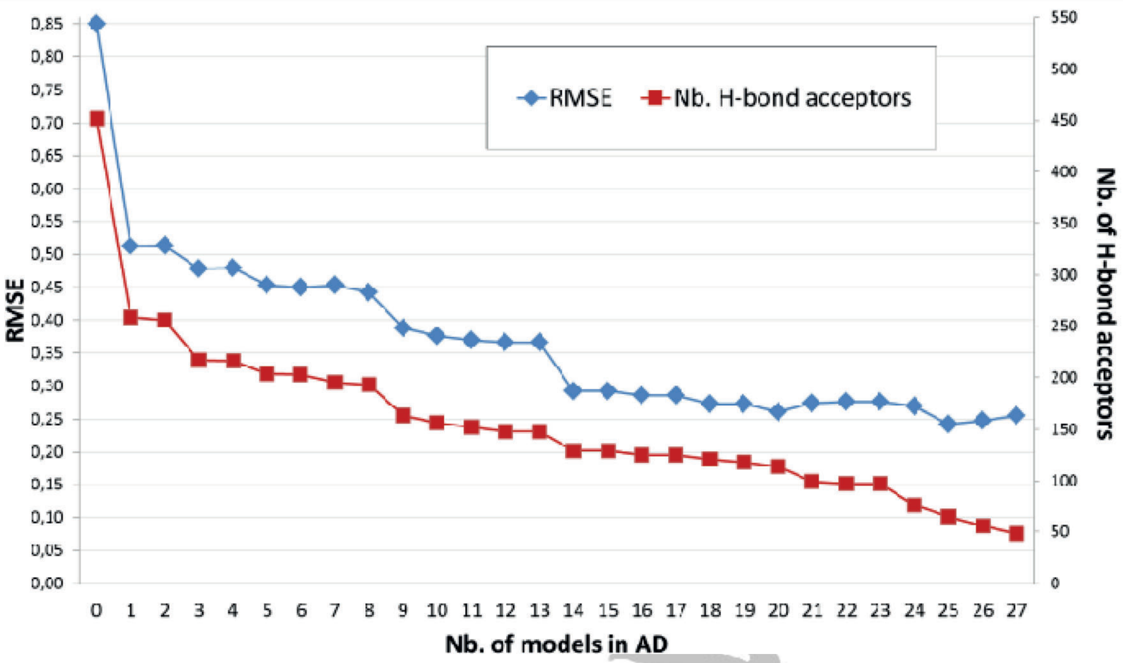


Figure 5. *RMSE* (blue diamond-shaped dots) and the number of H-bond acceptor centres (mono- and bifunctional) accepted by AD (red squares) by the SVM consensus model on the test set as a function of minimal number of applicable individual models

of accepted acceptors sites by AD as a function of the minimal number of applicable individual models (N_{mod}). At $N_{mod}=0$, the models' AD was not applied and predictions were made for all 451 acceptors sites. If, at least, one applicable model is applied ($N_{mod}=1$), the *RMSE* drastically drops because 40% of the test compounds are discarded. The increase of N_{mod} till 27 leads to further reduction of both *RMSE* (till 0.25) and the number of predicted acceptors sites (48). One can see that a good trade-off between performances and number of predictions is found at 14–15 applicable models, corresponding roughly to half of the entire ensemble of 27 models. This study shows the importance of considering AD which effectively discards the test set acceptor sites too dissimilar compared to the training set.

3.3 Bifunctional Molecules

Five bifunctional molecules are found in the training set and 47 in the test dataset. For the training set compounds, performance of the models was assessed in cross-validation. Without surprise, MOE and MA0 descriptors were not able to distinguish different acceptor sites of the same molecules, and therefore, they could not represent the bifunctional cases. It should be noted that MOE models correctly predicted one of the two sites. On the other hand, models based on MA2 or MA3 descriptors did not only provide individual predictions for different binding sites, but correctly ordered them (see Table 4 and supporting information Table SI 2).

Among the 47 bifunctional molecules of the test set (94 acceptor sites), 20 molecules contain acceptor types not present in the training set: aromatic or unsaturated moieties (forming H···π-bonds) and halogens (forming H···Hal bonds). All these irrelevant sites were removed by AD. The 27 remaining bifunctional molecules were considered to evaluate the models' performances.

For these 27 bifunctional compounds, the SVM CM with AD where a minimum of 14 models is required in CM leads to only 14 predicted values out of 54. The predictive performance (*RMSE*$=$0.25 and $R^2=0.85$) is statistically compatible with that observed in cross-validation and the overall prediction of the test set. The poor coverage of the model was expected since there were very few bifunctional compounds in the training set. The point is that, the substructural signature of most bifunctional species was never observed during training and those compounds were considered as out of AD. However, if the molecules are predicted when they are within the AD of at least one model, the *RMSE* is 0.44 which is still reasonable (see Supporting Information Section 2). Also note that these molecules were all indicated in the database as "approximated data". It can thus be assumed that the experimental error is greater on those molecules. The most pronounced outliers (see Figure 6) could be explained by either a small number of models applied for predictions or by inexact experimental data. For the most outlying molecule (indicated as 1 in Figure 6), experimental data has been measured with 4-nitrophenol[33] instead of 4-fluorophenol and then converted into pK_{BHX} by means of a LFER by Abraham et al.[59] Notice that this molecule can form an intra-molecular H-bond.

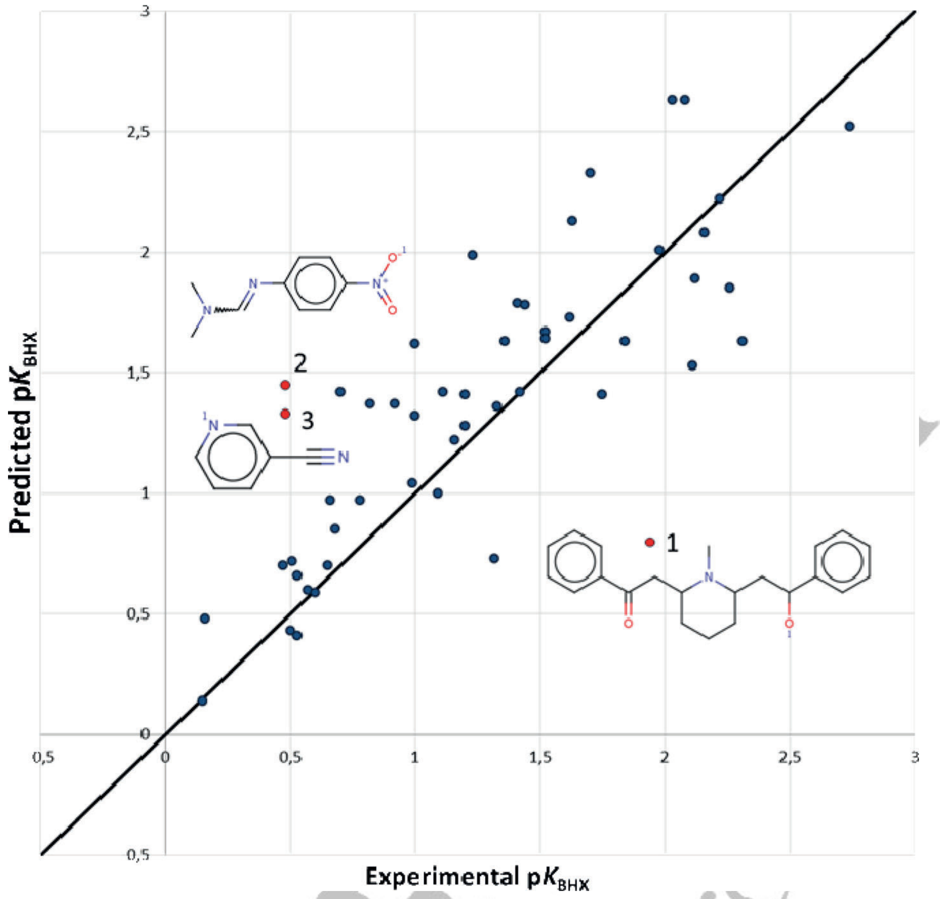


Figure 6. Prediction of 25 bifunctional molecules (51 values) by the SVM CM with the bounding box and fragment control as AD. Notice that three pK_{BHX} values for three acceptor sites were not predicted. The three structures shown correspond to the biggest prediction errors found.

3.4 Comparison with Previously Reported Models

One can hardly compare the performance of our models with that of LFER models of Abraham and Raevsky because the latter were not properly validated on a reasonably large external test set. As far as models based on quantum mechanics calculations is concerned, the best published model[7] achieves a $RMSE=0.13$ (calculated from the data reported in the Literature[7]). This model has been built on nitrogen HBAs restricting its application to the limited class of nitrogen compounds, despite the large panel of chemical functional groups encountered and the consideration of polyfunctional nitrogen structures. Our models perform slightly less good in cross-validation ($RMSE=0.25$–0.27, Table 2) and on the external test sets ($RMSE=0.25$–0.29, Table 3), but they are able to treat diverse sets of HBAs including nitrogen, oxygen, carbon, halogens, sulfur bases and the polyfunctional species with different atomic acceptor sites. They are clearly much less time-consuming because generation of fragment descriptors is a very fast procedure unlike heavy quantum mechanics calculations.

3.5 On-Line Implementation of the Models

The SVM consensus model is freely available on our web server: http://infochim.u-strasbg.fr/webserv/VSEngine.html (for more details see Supporting Information section 1). Two different approaches are supported. One provides an automatic detection of acceptor sites which is performed following the intrinsic ChemAxon "acceptor" type in pharmacophore mapping (PMapper).[61] This allows the user to submit plain, unmarked molecular files. A synthetic trust criterion of the prediction is provided, taking into account various aspects such as the number of individual models including the compound in their AD and the standard deviation of values predicted by individual models. For example, this tool has been tested in a screening of a database containing 2470 drugs and reference compounds from the US Pharmacopeia.[62] 10215 acceptor centres were identified and ~1700 centres (some 17% of the whole database) where predicted within the fragment control AD. The relatively small prediction rate can be explained by the fact that studied pharmaceutical compounds are more complex than H-acceptor molecules present in the training set.

Alternatively, specification of expected acceptor sites may be left in charge of the user, which may submit marked molecular files to the model. This enables the user to predict hydrogen bonding acceptor strength of centres not being considered as such by ChemAxon's PMapper.

4 Conclusions

This study presents new marked atom strategies for ISIDA fragment descriptors. This development was motivated by the need to better represent the locality of the H-bonding interaction as reported in the pK_{BHX} database.

The individual and consensus models built on large and structurally diverse data set of 542 H-bond acceptors were intensively validated both in cross-validation procedure and on two external datasets. The resulting models perform well both in cross-validation ($RMSE = 0.25$–0.27) and on the external test sets ($RMSE = 0.25$–0.29). Besides previously reported results, we demonstrate that our models based on marked atom descriptors are able to predict pK_{BHX} values in bifunctional molecules containing two different H-bond atomic sites, and, in principle, could treat polyfunctional species.

The SVM consensus model is publically available on the server: http://infochim.u-strasbg.fr/webserv/VSEngine.html. Only an internet access and browser is required to execute the models; no software installation is needed.

Supporting Information

The supporting information consists of two files: one for structures and pKBHX and one for modelling details.

Acknowledgements

We thank the *Centre of High-Performance Computing* of the Informatics Department of the University of Strasbourg (France) for the computational facilities.

References

[1] G. Gilli, P. Gilli, *The Nature of the Hydrogen Bond: Outline of a Comprehensive Hydrogen Bond Theory*, Oxford University Press, Oxford, **2009**.

[2] E. Arunan, G. R. Desiraju, R. A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D. C. Clary, R. H. Crabtree, J. J. Dannenberg, P. Hobza, H. G. Kjaergaard, A. C. Legon, B. Mennucci, D. J. Nesbitt, *Pure Appl. Chem.* **2011**, *83*, 1619–1636.

[3] C. Bissantz, B. Kuhn, M. Stahl, *J. Med. Chem.* **2010**, *53*, 5061–5084.

[4] C. Laurence, K. A. Brameld, J. Graton, J.-Y. Le Questel, E. Renault, *J. Med. Chem.* **2009**, *52*, 4073–4086.

[5] J. Graton, M. Berthelot, J.-F. Gal, C. Laurence, J. Lebreton, J.-Y. Le Questel, P.-C. Maria, R. Robins, *J. Org. Chem.* **2003**, *68*, 8208–8221.

[6] M. H. Abraham, P. P. Duce, D. V. Prior, D. G. Barratt, J. J. Morris, P. J. J. Taylor, *Chem. Soc., Perkin Trans. 2* **1989**, *10*, 1355–1375.

[7] F. Besseau, J. Graton, M. Berthelot, *Chem. Eur. J.* **2008**, *14*, 10656–10669.

[8] O. Lamarche, J. A. Platts, *Chem. Eur. J.* **2002**, *8*, 457–466.

[9] B. Pullman, ■ book title?■ Wiley, New York, **1978**.

[10] P. A. Kollman, L. C. Allen, *J. Am. Chem. Soc.* **1971**, *93*, 4991–5000.

[11] P. A. Kollman, J. McKelvey, A. Johansson, S. Rothenberg, *J. Am. Chem. Soc.* **1975**, *5*, 955–965.

[12] A. E. Kerdawy, C. S. Tautermann, T. Clark, T. Fox, *J. Chem. Inf. Model.* **2013**, *53*, 3262–3272.

[13] R. W. Taft, D. Gurka, L. Joris, P. v. R. Schleyer, J. W. Rakshys, *J. Am. Chem. Soc.* **1969**, *91*, 4801–4808.

[14] M. H. Abraham, P. L. Grellier, D. V. Prior, R. W. Taft, J. J. Morris, P. J. Taylor, C. Laurence, M. Berthelot, R. M. Doherty, M. J. Kamlet, J.-L. M. Abboud, K. Sraidi, G. Guiheneuf, *J. Am. Chem. Soc.* **1988**, *110*, 8534–8536.

[15] O. A. Raevsky, V. Y. Grigor'ev, V. P. Solovev, *Khim. Farm. Zh. (Russ.)* **1989**, *23*, 1294–1300.

[16] O. A. Raevsky, *J. Phys. Org. Chem.* **1997**, *10*, 405–413.

[17] O. A. Raevsky, V. Y. Grigor'ev, D. B. Kireev, N. S. Zefirov, *Quant. Struct.–Act. Relat.* **1992**, *11*, 49–63.

[18] R. S. Drago, B. B. Wayland, *J. Am. Chem. Soc.* **1965**, *87*, 3571–3577.

[19] A. V. Iogansen, *Teor. Eksp. Khim.* **1971**, *7*, 302–311.

[20] V. A. Terent'ev, *Thermodynamics of Hydrogen Bond*, Saratov University, Kuibyshev, **1973**.

[21] A. D. Sherry, K. F. Purcell, *J. Phys. Chem.* **1970**, *74*, 3535–3543.

[22] M. K. Kroeger, R. S. Drago, *J. Am. Chem. Soc.* **1981**, *103*, 3250–3262.

[23] O. A. Raevsky, V. V. Avidon, V. P. Novikov, *Khim. Farm. Zh. (Rus.)* **1982**, *16*, 968–971.

[24] O. A. Raevsky, V. Y. Grigor'ev, V. P. Solov'ev, I. V. Martynov, *Dokl. Akad. Nauk SSSR (Russ.)* **1988**, *298*, 1166–1169.

[25] J.-Y. Le Questel, M. Berthelot, C. Laurence, *J. Chem. Soc., Perkin Trans. 2* **1997**, 2711–2717.

[26] M. Hennemann, T. Clark, *J. Mol. Model.* **2002**, *8*, 95–101.

[27] A. S. Özen, F. D. Proft, V. Aviyente, P. Geerlings, *J. Phys. Chem. A* **2006**, *110*, 5860–5868.

[28] A. J. Green, P. L. A. Popelier, *J. Chem. Inf. Model.* **2014**, *54*, 553–561.

[29] A. Klamt, J. Reinisch, F. Eckert, J. Graton, J.-Y. Le Questel, *PhysChemChemPhys* **2013**, *15*, 7147–7154.

[30] A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov'ev, *J. Comput. Aid. Mol. Des.* **2005**, *19*, 693–703.

[31] J.-Y. Le Questel, G. Boquet, M. Berthelot, C. Laurence, *J. Phys. Chem. B* **2000**, *104*, 11816–11826.

[32] V. Arnaud, J.-Y. Le Questel, M. Mathé-Allainmat, J. Lebreton, M. Berthelot, *J. Phys. Chem. A* **2004**, *108*, 10740–10748.

[33] A. Locati, M. Berthelot, M. Evain, J. Lebreton, J.-Y. Le Questel, M. Mathé-Allainmat, A. Planchat, E. Renault, J. Graton, *J. Phys. Chem. A* **2007**, *111*, 6397–6405.

[34] V. Arnaud, M. Berthelot, F.-X. Felpin, J. Lebreton, J.-Y. Le Questel, J. Graton, *Eur. J. Org. Chem.* **2009**, 4939–4948.

[35] A. P. Atkinson, E. Baguet, N. Galland, J.-Y. Le Questel, A. Planchat, J. Graton, *Chem. Eur. J.* **2011**, 11637–11649.

[36] *JChem* Version 5.3.8, Calculator Plugin, ChemAxon, **2012**.

[37] V. Solovev, A. Varnek, *SDF manager EdiSDF* (Editor of Structure-Data Files), **2013**.

[38] O. A. Raevsky, V. P. Solov'ev, V. Y. Grigorev, *VINITI Depos. N 1001-V88*, Moscow, **1988**, p. 83.

[39] O. A. Raevsky, A. F. Solotnov, V. P. Solov'ev, *Zhurnal Obshchei Khimii (Rus)* **1987**, *57*, 1240–1243.

[40] A. Varnek, D. Fourches, D. Horvath, O. Klimchuk, S. Gaudin, P. Vayer, V. P. Solovev, F. Hoonakker, I. V. Tetko, G. Marcou, *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 191 – 198.

[41] F. Ruggiu, G. Marcou, A. Varnek, D. Horvath, *Mol. Inf.* **2010**, *29*, 855 – 868.

[42] V. P. Solovev, A. Varnek, G. Wipff, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 847 – 858.

[43] *MOE*, Chemical Computing Group Inc., Montreal, **2011**.

[44] A. T. Hagler, E. Huler, S. Lifson, *J. Am. Chem. Soc.* **1974**, *96*, 5319 – 5327.

[45] *ISIDA Fragmentor2012*, Laboratoire de Chémoinformatique, UMR 7140, Université de Strasbourg, France, **2012**.

[46] V. P. Solovev, A. Varnek, ISIDA (*In Silico Design and Data Analysis*) *QSPR Program*, v5.76, Strasbourg – Moscow, **2012**.

[47] L. Xing, R. C. Glen, R. D. Clark, *J. Chem. Inf. Model.* **2003**, *43*, 870 – 879.

[48] S. Jelfs, P. Ertl, P. Selzer, *J. Chem. Inf. Model.* **2007**, *47*, 450 – 459.

[49] M. Rupp, R. Körner, I. V. Tetko, *Mol. Inf.* **2010**, *29*, 731 – 740.

[50] C. C. Chang, C.-J. Lin, *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1 – 27.

[51] T. G. Dietterich, *Neural Comput.* **1998**, *7*, 1895 – 1923.

[52] K. M. Ali, M. J. Pazzani, *Machine Learning* **1996**, *24*, 173 – 202.

[53] D. Horvath, G. Marcou, A. Varnek, *J. Chem. Inf. Model.* **2009**, *49*, 1762 – 1776.

[54] I. V. Tetko, I. Sushko, A. K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Öberg, R. Todeschini, D. Fourches, A. Varnek, *J. Chem. Inf. Model.* **2008**, *48*, 1733 – 1746.

[55] C. Rücker, G. Rücker, M. Meringer, *J. Chem. Inf. Model.* **2007**, *47*, 2345 – 2357.

[56] A. Varnek, N. Kireeva, I. V. Tetko, I. Baskin, V. P. Solovev, *J. Chem. Inf. Model.* **2007**, *47*, 1111 – 1122.

[57] V. P. Solov'ev, A. Varnek, *Rus. Chem. Bull.* **2004**, *53*, 1434 – 1445.

[58] V. P. Solov'ev, A. Y. Tsivadze, A. Varnek, *Macroheterocycles* **2012**, *5*, 404 – 410.

[59] M. H. Abraham, P. L. Grellier, D. V. Prior, J. J. Morris, P. J. Taylor, *J. Chem. Soc. Perkin Trans. 2* **1990**, *0*, 521 – 529.

[60] E. Arunan, G. R. Desiraju, R. A. Klein, J. Sadlej, S. Scheiner, I. Alkorta, D. C. Clary, R. H. Crabtree, J. J. Dannenberg, P. Hobza, H. G. Kjaergaard, A. C. Legon, B. Mennucci, D. J. Nesbitt, *Pure Appl. Chem.* **2011**, *83*, 1637 – 1641.

[61] *ChemAxon PMapper*, JChem v. 6.1.7 https://www.chemaxon.com/jchem/doc/user/PMapper.html, **2014.**

[62] *United States Pharmacopeia*, http://www.usp.org/, **2014**.

Received: March 12, 2014
Accepted: May 15, 2014
Published online: ■■ ▮▮, 0000