analytical chemistry

Article

pubs.acs.org/ac

Quantitative Structure–Property Relationship Modeling: A Valuable Support in High-Throughput Screening Quality Control

Fiorella Ruggiu,[†,‡] Patrick Gizzi,[§,▽,‡] Jean-Luc Galzi,[§,▽] Marcel Hibert,[∥,▽] Jacques Haiech,[∥,▽] Igor Baskin,[†,⊥] Dragos Horvath,[†] Gilles Marcou,[†] and Alexandre Varnek*[,†]

[†]Laboratoire de Chémoinformatique, UMR 7140 CNRS, Université de Strasbourg, 1 rue Blaise Pascal, 67000 Strasbourg, France

[§]Laboratoire de Biotechnologie et Signalisation Cellulaire (Plate-forme TechMedILL), UMR 7242 CNRS, Ecole Supérieure de Biotechnologie Strasbourg, Université de Strasbourg, 67412 Illkirch Graffenstaden, France

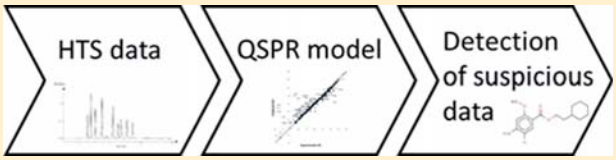
[∥]Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS, Faculté de Pharmacie, Université de Strasbourg, 74 route du Rhin, 67401 Illkirch, France

[⊥]Lomonosov Moscow State University, Moscow 119991, Russia

[▽]Laboratory of Excellence Medalis, IBMC du CNRS, Université de Strasbourg, 15 rue René Descartes, 67000 Strasbourg, France

Supporting Information

ABSTRACT: Evaluation of important pharmacokinetic properties such as hydrophobicity by high-throughput screening (HTS) methods is a major issue in drug discovery. In this paper, we present measurements of the chromatographic hydrophobicity index (CHI) on a subset of the French chemical library Chimiothèque Nationale (CN). The data were used in quantitative structure–property relationship (QSPR) modeling in order to annotate the CN. An algorithm is proposed to detect problematic molecules with large prediction errors, called outliers. In order to find an explanation for these large discrepancies between predicted and experimental values, these compounds were reanalyzed experimentally. As the first selected outliers indeed had experimental problems, including hydrolysis or sheer absence of expected structure, we herewith propose the use of QSPR as a support tool for quality control of screening data and encourage cooperation between experimental and theoretical teams to improve results. The corrected data were used to produce a model, which is freely available on our web server at http://infochim.u-strasbg.fr/webserv/VSEngine.html.



Since the advent of robotized biological testing in the 1990s, access to large, diverse, and original compound collections has become a major issue in drug discovery. However, handling of such collections raises important logistical and technical challenges, in particular because compound originality, a prerequisite for patentability, is by definition not the hallmark of standard, well-conditioned commercial collections accessible to everyone. Extensive analytical assessment of purchased compound collections is therefore a time-consuming and cost-intensive key issue, for its automation may go only as far as automated recording followed by error-prone machine interpretation of analysis results. Time and resources for in-depth structural analysis is lacking; therefore, standard purity measures are necessary but hardly sufficient.[1,2] In standard liquid chromatography/mass spectrometry (LC/MS) analysis, purity is taken as granted if an LC peak of expected mass is "predominant". However, the tacit assumptions that (a) the correct mass actually stands for the expected isomer and (b) the sensitivity of the detector is the same for the main compound and the potential impurities are virtually never checked. In practice, in-depth structural analysis is postponed to the hit reconfirmation stage, for allegedly active molecules only.

In this context, academic compound collections such as the Chimiothèque Nationale (CN), the French national chemical library regrouping original compounds issued from nationwide academic research, is a valuable asset in terms of originality and diversity but a logistical nightmare. Compounds are issued from different laboratories, conditioned according to different operating rules, and stored under variable conditions before being sent to the central repository. The CN therefore requires quality control. A "Projet Interdisciplinaire de Recherche" (PIR) has been conceived as a showcase project to illustrate the use of this collection in (high) throughput screening (HTS) tests and to highlight and fix various pitfalls due to the peculiar nature of this collection. PIR was aimed at annotating the CN with respect to hydrophobicity, solubility, and acidity by using a diverse subset of 640 molecules, named the "Chimiothèque Nationale Essentielle" (CNE), as a representative core of the CN. It was not tailored for drug design and therefore includes reactive and nondruglike molecules as well. The CNE molecules were then cherry-picked and submitted to standard

Received: November 2, 2013
Accepted: January 30, 2014
Published: January 30, 2014

dx.doi.org/10.1021/ac403544k | *Anal. Chem.* 2014, 86, 2510–2520

quality control (QC) based on LC/MS purity check at the Integrative Chemical Biology Platform of Strasbourg (PCBIS).

Parallelized and rapid measuring of the envisaged physicochemical properties was carried out at the TechMed^{ILL} Platform in Strasbourg. Hydrophobicity—the first measured property and the one concerned in this paper—is an important property for medicinal chemists.[3] It is widely used as a criterion for acceptable drug solubility and permeability.[4] It has been shown to be related to absorption distribution metabolism excretion/toxicity (ADME/T) properties for over a century.[5] It has classically been evaluated by the octanol−water partition coefficient log $P_{o/w}$ after the proposal of Hansch and Fujita[6] and measured by the shake-flask method. However, this method is time-consuming and a modern HTS method using high-pressure liquid chromatography (HPLC) originally developed by GlaxoSmithKline researchers[7,8] has been used to assess the CNE, the chromatographic hydrophobicity index (CHI).

In reverse-phase HPLC, the partition between a hydro-organic mobile phase and a C-18 stationary phase is governed by hydrophobicity. The organic solvent percentage in mobile phase necessary for elution is referred to as the isocratic chromatographic hydrophobicity index (ICHI), which is thus a good alternative to log $P_{o/w}$ measures.[9] However, this measure requires testing several mobile phases with different organic solvent percentages and thus is time- and resource-consuming. This is why an alternative method based on a fast gradient was developed. The measured retention time in such columns are linearly correlated to ICHI[7] and to log $P_{o/w}$.[8] The method uses a linear calibration generated from the retention times obtained for a set of 10 standard compounds with known ICHI values. For any new compound, the retention time transformed by this calibration gives a number which is referred to as the CHI. This method is cost-effective and very economical in terms of compound requirement and solvent. To conclude, CHI is a measure of retention of the test compound on a fast gradient C-18 column.

It shall be noted that for compounds whose retention is not significant, a negative CHI value will be returned, meaning very low hydrophobicity. For compounds that are not easily washed off the column, a CHI value of >100 is obtained, signifying very high hydrophobicity. But a linear relationship between CHI and ICHI is observed only between 18.4 and 96.4 (the most extreme calibration values). It is important to note that this CHI range covers that of molecules that cross intestinal and brain barriers spontaneously. Molecules with CHI <0 or >100 are not useful in drug discovery programs.

Chemoinformaticians exploited the measured CHI data to build associated quantitative structure−property relationship (QSPR) models on the basis of the CNE diverse training set. The aim was to build useful models in order to annotate all the other academic molecules of the CN by their predicted properties and also to enable chemists to make predictions for novel structures, via a publicly accessible QSPR prediction web server. QSPR models are mathematical models fitted on the data that return an estimate of the expected property on the basis of molecular descriptors serving to numerically encode the features present in the chemical structure. Parameter fitting is done to ensure that, for each training compound (of known property Y), the model will return a predicted Y_{pred} very close to Y (following the classical least-squares principle). The molecular descriptors used in this study are the ISIDA property-labeled fragment counts.[10] Fitting was performed mainly by use of support vector machines (SVM),[11] because of the robustness of the produced models. Other machine learning methods were also tried out.

The main insights gained from this work come from the systematic failures observed in modeling. We define *outliers* as compounds for which their calculated property value Y_{pred} could never be brought in agreement with the observed Y, irrespectively of the employed model-building strategy. This is in line with the classical definition of an outlier as an observation that is numerically distant from the rest of the data.[12] We propose a method for their systematic annotation and then to submit them to in-depth experimental scrutiny. The observed discrepancies between Y and Y_{pred} were much higher than the expected model imprecisions, and yet independent of modeling premises it was hypothesized that this could be due to real differences in molecular structures: thus the actual molecule returning the measured Y might not correspond to the nominal structure for which Y_{pred} was estimated. We identified three periods during which a chemical alteration might have occurred: (a) since the CNE QC, during storage; (b) before the CNE QC, without being detected at that stage; or (c) during the actual hydrophobicity measurement, due to reaction with the aqueous buffer.

Systematic analysis of outliers actually revealed the above hypothesis to be basically correct. This signifies that a properly built QSPR model (with minimized modeling artifacts such as overfitting) is robust enough to highlight experimental errors. Building a QSPR model in parallel to experimental assessment of a library is not a costly undertaking and may effectively pinpoint potential experimental pitfalls, focusing the need for in-depth further analysis to the potentially "pathological" items. This could be an important first step toward the use of QSPR approaches for regulatory purposes, instead of experimental measurements, as envisaged by the REACH project (for registration, evaluation, authorization, and restriction of chemicals).[13]

This paper is organized in order to follow the chronology of the different experimental and modeling steps within the study. First the experimental protocol and results of the CHI measurements is presented, followed by an outline of the computational procedures and the outlier management section. Outlier management contains the initial building of the models, the modeling protocol for the identification of the outliers, their experimental validation, and a presentation of the results with a discussion. Finally, the consensus model, build after removal of outliers and doubtful molecules from the set is presented, followed by a conclusion section.

CHROMATOGRAPHIC HYDROPHOBICITY INDEX MEASUREMENTS

The 640 CNE compounds were received in eight microplates containing 10 mM dimethyl sulfoxide (DMSO) stock solutions. CHI measurements were done on a Gilson HPLC system with a photodiode array detector, an autosampler, and a Valco injector. Data acquisition and processing were performed with Trilution LC V2.0 software. Measurements were carried out at 20 ± 2 °C. A 5 μm Luna C18(2) column (50 × 4.6) purchased from Phenomenex was used. The mobile phase flow rate was 2 mL/min and the following program was applied for the elution: 0−0.2 min, 0% B; 0.2−2.7 min, 0−100% B; 2.7−3.2 min, 100% B; 3.2−3.4 min, 100−0% B; and 3.4−6.1 min, 0% B. Solvent A was 50 mM ammonium acetate (pH 7.4) in water, and solvent B was HPLC-grade acetonitrile (Sigma−Aldrich Chromasolv). The detection wavelengths were 254 and 230 nm.

dx.doi.org/10.1021/ac403544k | Anal. Chem. 2014, 86, 2510−2520

First, a solution with 10 reference compounds with known ICHI values (see Supporting Information section 1) was injected onto the HPLC to generate a calibration line from their retention times (see Figure 1). The concentration of the

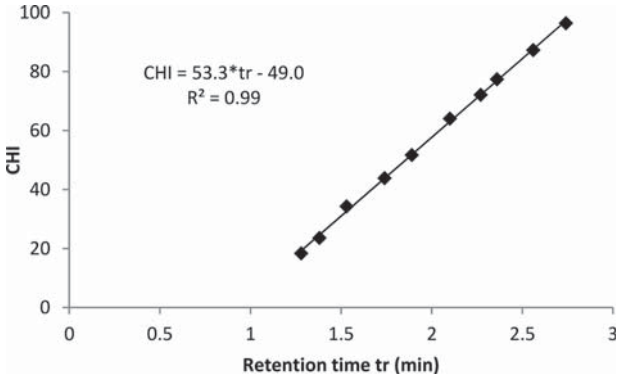


Figure 1. Calibration of the HPLC column: relationship between retention times and CHI values.

mixture was 0.2 mg/mL for each compound and the injected volume was 3 μL. A typical chromatogram of the standard solution is represented in Figure 2. The test compounds were analyzed on the same system. The 10 mM DMSO stock solutions were diluted to 200 μM in acetonitrile/50 mM ammonium acetate, pH 7.4 (1/1 v/v). The linear regression equation of the calibration line was used to convert retention time of the test compounds to CHI values (CHI 1 in Table 1).

The experimental procedure for CHI measurement was applied to all 640 molecules of CNE, and several experimental complications arose (see Figure 3). CHI values of 418 compounds were measured without any complications. The protocol is based on ultraviolet–visible (UV–vis) detection; therefore, compounds lacking chromophore moieties cannot be detected by this method, which was the case for 10% of the molecules. In addition, nothing has been detected for 4% of the molecules for unknown and probably undefined reasons (presumably compound insolubility or instability in DMSO or degradation in test buffer). Several peaks were detected for 36 compounds (6%), indicating impurity or degradation. Hence, matching a peak to the molecule drawn in the database is difficult. It was assumed that the most intense peak corresponds to it. Compounds that gave peaks with low intensity were considered but with caution, because it demonstrates a solubility problem. Finally, CHI values were measured for 545 molecules and complications were annotated in the database.

COMPUTATIONAL PROCEDURE

The computational workflow used in this work is given in Figure 4. Steps 1–5 are described in this section, whereas steps 6–8 are reported under Final Consensus Model.

Compound Standardization. The molecules were standardized by removing salts, stripping off hydrogens from the molecular graph, choosing a standard representation for groups such as nitro or imidazole, and generating major tautomer as well as major microspecies at pH = 7.4 with ChemAxon's Calculator plugin.[14]

Descriptors Calculation. ISIDA property-labeled descriptors,[10] a type of fragment count descriptors, were calculated. Sequences, extended augmented atoms, and triplets were computed on the molecular graph, which has been "colored" with one of the following properties: atomic symbols, pharmacophoric flagging, electrostatic potentials, or force field typing. The length of fragments varied for the minimum from 2 to 4 and for the maximum from 4 to 8. Further variants were then introduced for some of these, by toggling additional options: switching to "Atom pairs" mode, enabling "all path exploration", and the explicit representation of the formal charge. A total of 2772 descriptor pools were eventually generated.

Machine Learning Techniques. SVM was chosen as the reference machine learning because of its stability, mainly due to its particular error function. The Libsvm 3.12 package[11] was used for generation of ε-SVM regression models with a linear kernel, and ε was set equal to the random experimental error estimated at 2 CHI units. The cost was tested for 28 different values ranging from 0.1 to 100. Model building included both operational parameters fitting (as required by the libsvm approach) and, most important, required cross-validation

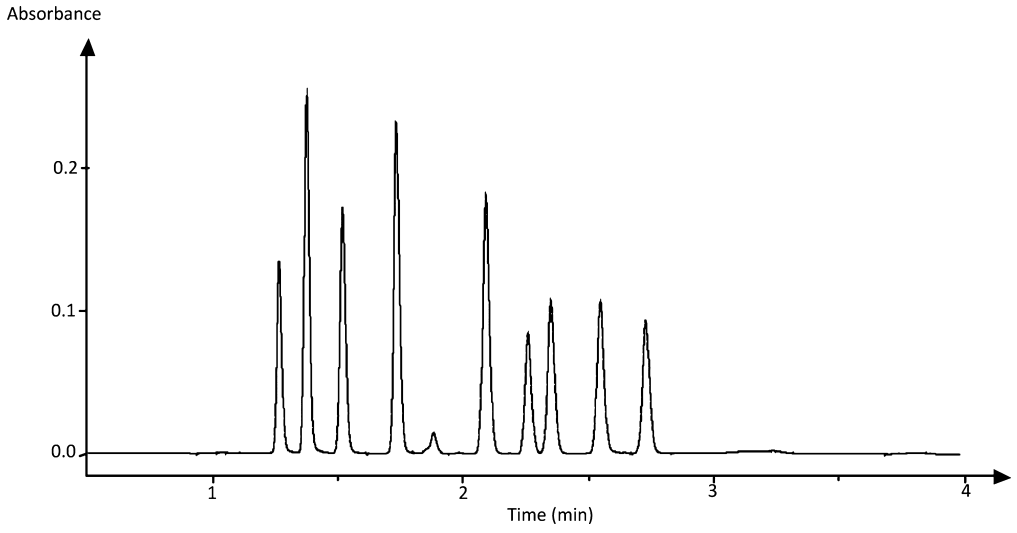


Figure 2. Typical chromatogram of the standard solution.

dx.doi.org/10.1021/ac403544k | Anal. Chem. 2014, 86, 2510–2520

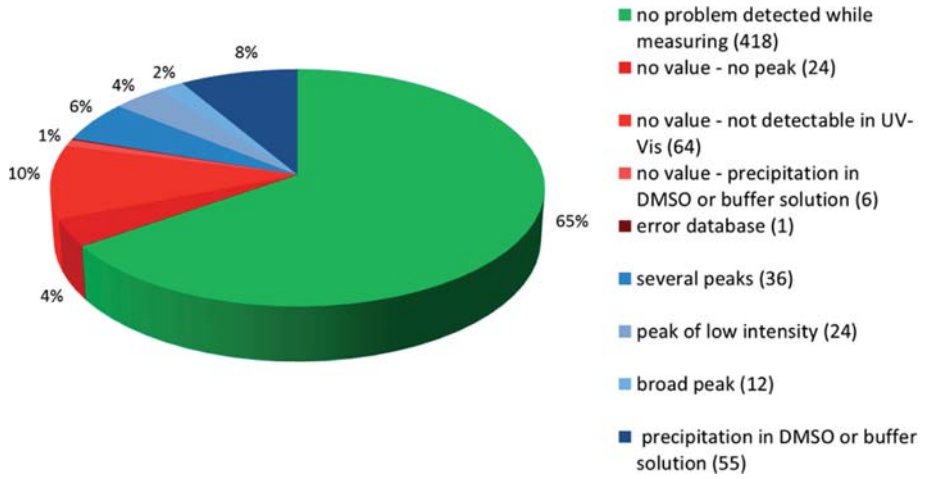


Figure 3. Experimental status of CHI measurements on 640 molecules: green, no problems detected; red, failures to determine the CHI value; and blue, measurements accompanied by observed side phenomena that may signal artifacts, all while nevertheless allowing some CHI value to be recorded.

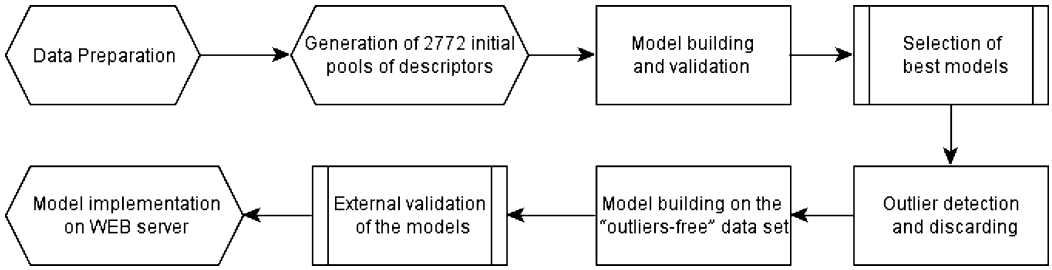


Figure 4. Computational workflow used in this work.

techniques[15] to avoid overfitting. The final model selection criterion therefore was 5-fold cross-validated root-mean-squared error (5CV-RMSE) (see Supporting Information section 2 for details on statistical parameters).

Partial least-squares (PLS) regression models[16] and stochastic quantitative structure–activity relationship (QSAR) sampler (SQS) regression models[17] issued from selected pools of descriptors were also built for comparison purposes.

Model Selection. In total, 2772 × 28 = 77 616 individual models (each corresponding to a particular descriptor pool and a particular value of cost parameter) have been obtained for a given data set. Several "best" models were selected according to 5CV-RMSE. All selected models were used for consensus predictions on the external test set: for each molecule, the CHI value was calculated as an arithmetic average of predictions made by selected individual models.

Outlier Identification Protocol. In this section we discuss the identification of recurrent outliers observed in different modeling strategies. The term "outlier" designates, in the following, a compound for which the predicted value returned by a model having used this molecule for learning strongly diverges from the experimental value.

The list of outliers, submitted to in-depth analysis in order to attempt reconfirmation of these experimental values that could not be explained by modeling, was gathered by an *eliminate-and-refit* protocol on the basis of N best models. At each step of the prediction, a given data point is considered anomalous if its calculation error at the fitting stage is higher than a threshold C_{out}. This threshold is computed as twice the highest 5CV-RMSE found in the set of N values from each SVM model: C_{out} = 2max(5CV-RMSE). The outlier list was iteratively built as follows:

(1) The molecule with the highest number of anomalous estimates is chosen, based on the current value of C_{out}. In the event of a tie, the molecule with the highest absolute mean prediction error is chosen.

(2) The corresponding compound is removed from the modeling data set and the N models are refitted. The operational parameters are not reoptimized.

(3) The experimentally measured CHI value in discrepancy with the prediction is challenged, by a thorough reanalysis of the compound (see Experimental Reassessment of Outliers).

(4) The procedure is repeated from step 1 until no more of the apparently irreconcilable experiment–prediction discrepancies can be attributed to measurement problems (cases a–c listed previously).

The choice of using fitted values is more logical than using 5CV-predicted values as model "output" to compare to the experimental value. Indeed, discrepancies between 5-CV-predicted values and experiment are more likely to occur, especially for species at the edge or outside the applicability domain.[18] If the model has already learned from a molecule, it should be able to predict it. However, if the fitted value of a molecule is in discrepancy with the measured data, this indicates that the molecule goes against what the model learned from other molecules. The stepwise manner of this protocol for picking out outliers instead of selecting several on the same model ensures that the presence of the biggest outlier does not

dx.doi.org/10.1021/ac403544k | *Anal. Chem.* 2014, 86, 2510–2520

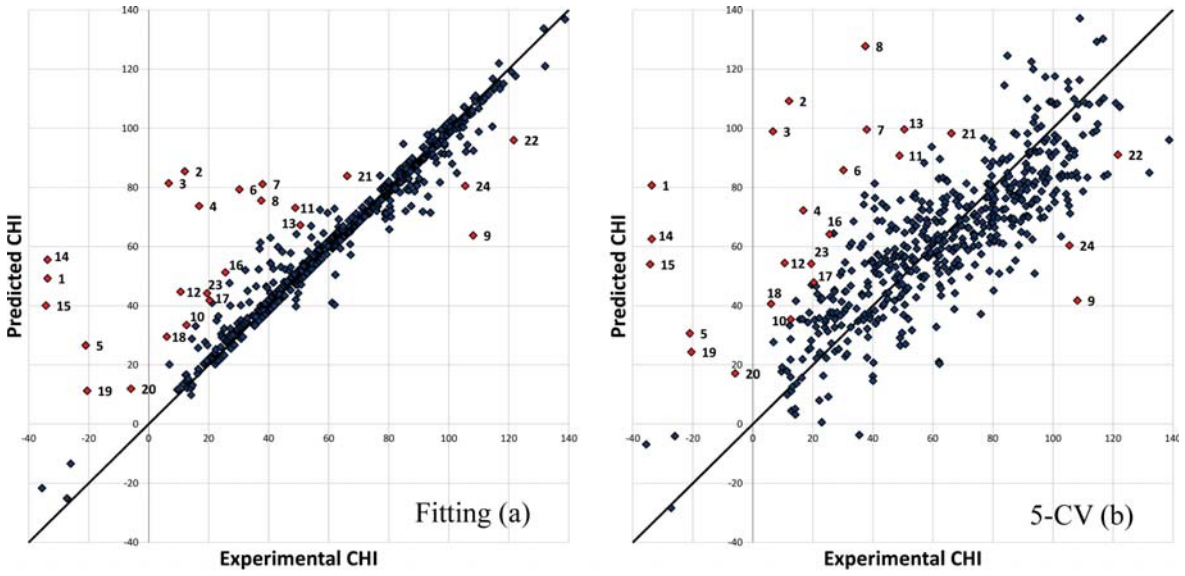


Figure 5. Experimental vs predicted CHI assessed (a) at the fitting stage and (b) in 5-fold cross-validation for the best SVM model (see Outlier Detection, Validation, and Analysis). The numbers indicate the outliers detected in the *eliminate-and-refit* protocol and listed in Table 1.

significantly skew the calculated values for other compounds. When one molecule is eliminated from the training set, the model is refitted and changes. Thus, it cannot be assumed that the molecule with the biggest error on the rebuild model is the same as the second biggest in the initial model. Besides, the fact that a compound appears as outlier for several models is a concept of paramount importance to this analysis because it permits convergence toward problematic molecules identified by different points of views.

OUTLIER DETECTION, VALIDATION, AND ANALYSIS

Outlier Detection. Ten models out of 77 616 built on the parent set of 545 compounds were selected according to 5CV-RMSE. The best of them involves atom-centric fragments colored by atomic symbols with a range of 2−4 atoms, with the use of formal charges, and with a SVM cost of 0.5. It has a train-RMSE of 11.2 and a 5CV-RMSE of 19.6. The obtained models show several recurrent outliers (see Figure 5).

The CNE set is the biggest collection of CHI values found in the literature. It is a very reliable source of data, as it was measured by the same scientist, with the same equipment, in the same conditions (room temperature, solutions used). Thus, the hypothesis that the data cannot be modeled due to multiple protocol incoherencies was discarded. A closer analysis of the structure of those molecules showed that some contained potentially reactive groups, leading us to foresee that problems may concern certain experimentally measured values, even though, in most cases, no peculiar complications were noted during these measurements.

In order to check if the relatively poor model performance is due to inclusion in the training set of molecules for which some experimental complications were detected (blue portion of the chart in Figure 3), modeling was performed on the set of 418 molecules measured without any complications (green portion of the chart in Figure 3). We did not observe any significant improvement of performance, and thus it was expected that reported experimental problems were not indicative of data limiting the quality of the models, as outliers would.

If experimental annotation were not sufficient to discard suspicious data, the question was to which extent are QSPR models able to highlight problems in a set of data issued from an HTS experiment? On the one hand, it is interesting to see how many of those with known experimental problems are perceived as outliers. Are outliers with no apparent experimental problems affected by issues that were not observable during the CHI measurement protocol?

To answer these questions, the *eliminate-and-refit* protocol described under Computational Procedure has been applied for the 10 best SVM models (see the model parameters in Supporting Information section 3). This led to the detection of the 24 outliers listed in Table 1. Unsurprisingly, outliers detected at the fitting stage also behave erratically during 5CV (see Figure 5).

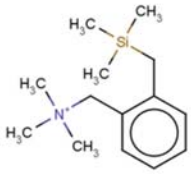
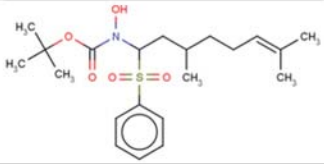
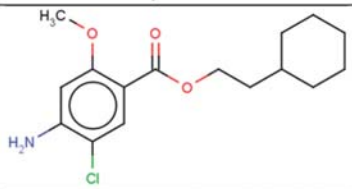
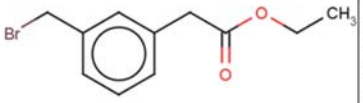
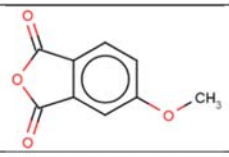
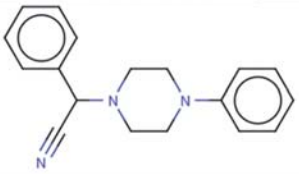
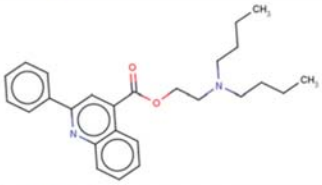
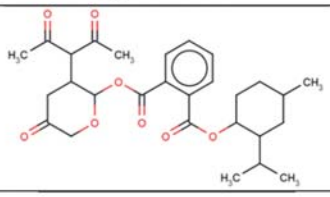
To ensure the outliers did not contain unique features that would make them fundamentally different from the others in the training, 1-SVM[19] using a linear kernel was applied at varying ν parameters. The outlier distribution is homogeneous within the data set. The percentage coverage within the outliers corresponds to the percentage coverage within the data set. If these outliers differed structurally from the other molecules within the set, they would never be within the dense area defined by the 1-SVM.

Experimental Reassessment of Outliers. The experimental check of compounds annotated as outliers was done by the TechMed^{ILL} Plateform. CHI values of the compounds identified as outliers were measured a second time (CHI 2 in Table 1) and solutions were submitted to mass spectrometry recharacterization in order to explain differences found between experimental and predicted CHI values. Fresh DMSO stock solutions were prepared from powders except for four compounds for which powder was not available (indicated by asterisks in Table 1). The powder should contain less impurities and eventual chemical degradation is less likely to occur than in the stock solution.

First, these solutions were used to determine the CHI values again by the same procedure explained previously (see Chromatographic Hydrophobicity Index Measurements),

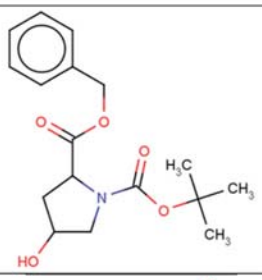
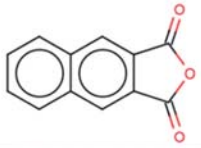
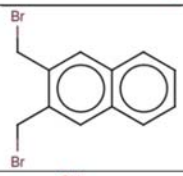
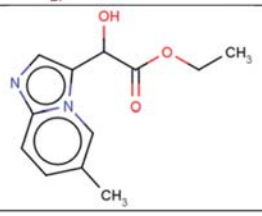
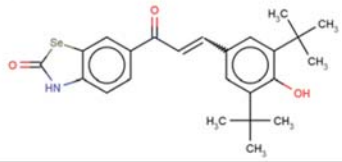
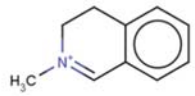
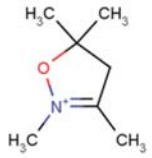
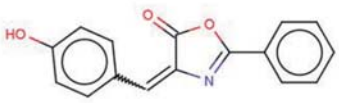
dx.doi.org/10.1021/ac403544k | *Anal. Chem.* 2014, 86, 2510−2520

Table 1. Outliers List and Experimental Results[a]

Nb	Theoretical Structure	Comments	CHI 1	CHI_{pred}	CHI 2	MS
1		Desired compound presence is confirmed by MS but this product is not detected by UV. Indeed, for both CHI measurements, only one peak is detected at the void time, which corresponds to a CHI value of -34. The compound not retained by the column is not identified.	-33.7	56.4	-33.7	Y
2		The desired compound is not observed by MS. It is probably insoluble in the buffer. The UV peaks detected for CHI measurements refer to an unknown product.	12.1	88.5	9.6	N
3		The acid resulting from the hydrolysis of the ester is detected by MS. The low value obtained for CHI1 experiment is explained by this hydrolysis. The second CHI measurement with a fresh solution allows detecting the expected ester.	6.7	82.2	108.9	Y
4		The well used for CHI1 measurement contains the desired compound but at a very low concentration confirmed by a small MS response and not detectable by UV. A contaminant with a low hydrophobicity is observed by MS. CHI2 experiment allows detecting the desired compound.	16.8	76.5	86.4 and 80	Y
5*		Although the desired compound is detected by MS with a small response, the major product in the well is the diacid resulting from the hydrolysis of the anhydride.	-21.0	38.4	-24.6	Y
6		The desired compound is detected by MS but as a minor product in the solution used for CHI1 experiment. A contaminant is found. The second measurement CHI2 allows detecting the expected compound as the major product.	30.2	84.7	99.8 and smaller peak at 34.1	Y
7		The desired compound is detected by MS but as a minor product. The UV peaks detected for CHI measurements refer to an unknown product.	38.0	89.4	36.8	Y
8		The desired compound is detected by MS with a very small response. The corresponding concentration is probably not detectable by UV. Two other products are observed. The diacid resulting from the hydrolysis of the esters is detected.	37.5	87.9	33.6 and 58.7	Y

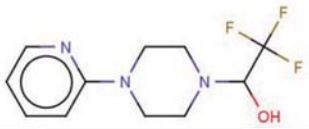
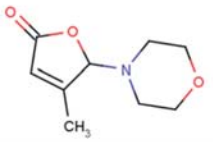
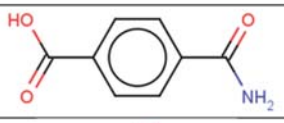
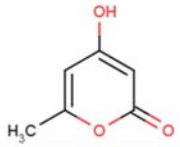
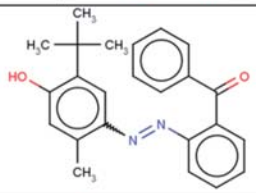
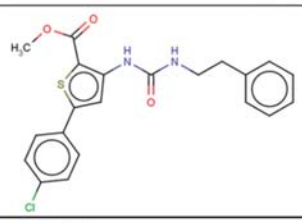
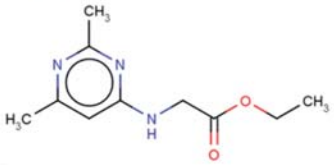
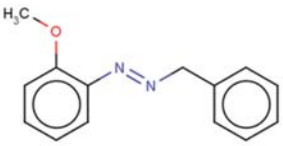
dx.doi.org/10.1021/ac403544k | Anal. Chem. 2014, 86, 2510−2520

Table 1. continued

Nb	Theoretical Structure	Comments	CHI 1	CHI_{pred}	CHI 2	MS
9		The desired compound is detected by MS but as a minor product in the solution used for CHI1 experiment. A contaminant is found. The second measurement CHI2 allows detecting the expected compound as the major product.	108.2	67.5	71 and 101.9	Y
10*		Although the desired compound is detected by MS with a small response, the major product in the well is the diacid resulting from the hydrolysis of the acid anhydride.	12.6	53.4	11.2	Y
11		The desired compound is observed by MS but with a very low response. It is probably insoluble in the buffer. The UV peaks detected for CHI measurements refer to an unknown product.	48.9	91.9	49.1	Y
12		The desired compound is detected by MS but as a minor product in the solution used for CHI1 experiment. The acid resulting from the hydrolysis of the ester is detected by MS. The second measurement CHI2 allows detecting the expected compound as the major product.	10.6	47.1	10.1 and 42.1	Y
13		The desired compound's presence is confirmed by MS. The first and the second CHI measurements do not match.	50.5	89.84	114.4	Y
14		The desired compound's presence is confirmed by MS but not detected by UV. Indeed, for both CHI measurements, only one peak is detected at the void time, which corresponds to a CHI value of -34. This compound not retained by the column is not identified. CHI2 experiment allows detecting the desired compound.	-33.7	26.3	-33.7 and 20.2	Y
15*		The presence of the desired compound is confirmed by MS. As it does not contain any chromophore, it cannot be detected by UV. The peak detected at the void time for CHI measurements corresponds to a CHI value of -34. The compound not retained by the column is not identified.	-34.3	18.1	-33.7	Y
16		The desired compound is not observed and the acid resulting from the hydrolysis of the lactone is detected by MS. The low value obtained for CHI1 experiment is explained by this hydrolysis.	25.6	59.1	24.5	N

dx.doi.org/10.1021/ac403544k | *Anal. Chem.* 2014, 86, 2510−2520

Table 1. continued

Nb	Theoretical Structure	Comments	CHI 1	CHI_{pred}	CHI 2	MS
17		The desired compound is not detected by MS while the substructure without the $C(CF_3)OH$ is observed.	20.3	49.5	27.1	N
18		The presence of the desired compound is confirmed by MS. The first and the second CHI measurements do not match.	6.1	39.7	-28.9	Y
19		The desired compound is not detected by MS. Both CHI measurements give identical results but do not correspond to the expected product.	-20.5	12.0	-24.6	N
20*		Compound's presence is confirmed by MS but as it does not contain any chromophore, it cannot be detected by UV. The UV peak detected for CHI2 measurement refers to an unknown product.	-5.9	25.1	27.7	Y
21		The presence of the desired compound is confirmed by MS. The first and the second CHI measurement do not match.	66.2	97.7	121.65	Y
22		No problem detected.	121.7	94.0	116.3	Y
23		No problem detected.	19.5	47.1	22.4	Y
24		No problem detected.	105.5	80.0	101.4	Y

[a]CHI 1 is the first CHI value, obtained with DMSO solutions in plates received from the central repository (the whole set was measured with UV−vis detection and used for the first modeling). CHI_{pred} stands for CHI average prediction and corresponds to the average prediction over the 10 best SVM models in the iterative procedure. CHI 2 is the second CHI value, obtained with fresh solutions prepared from powders (except for those marked with asterisks) and measured for the 24 outliers (with LC/UV). MS indicates whether the presence of the theoretical structure was confirmed (Y) or invalidated (N) by mass spectrometry

which permits us to check whether the stock solutions distributed by the CN had problems. Second, a LC/MS characterization was done to confirm or invalidate the presence of the expected compound (see MS column in Table 1), as described by its theoretical structure in the database. Any error in this drawn structure will induce an error in the QSPR estimate, as the descriptors calculated will not correspond to the actual measured structure. A LCMS-8030 triple-quadrupole liquid chromatograph mass spectrometer was used for these quality control measurements. Ionization of compounds was done with an electrospray source. Both single-ion monitoring and scan modes were used. The first mode was applied in order

dx.doi.org/10.1021/ac403544k | Anal. Chem. 2014, 86, 2510−2520

to control whether the compounds in solution match with the given structures. The second mode allowed identification of other compounds present in the solution, such as impurities or products of degradation. As mass spectrometers do not support high flow rates and high salt concentration in mobile phase, it was impossible to reproduce the same experimental conditions of CHI measurements. Data acquisition and processing were performed with Labsolutions v5.0 software. Measurements were carried out at 25 °C. A 1.7 μm Kinetex C18 column (50 × 2.1) purchased from Phenomenex was used. The mobile phase flow rate was fixed at 0.5 mL/min and the following program was applied for the elution: 0−0.2 min, 0% B; 0.2−3 min, 0−100% B; 3−3.2 min, 100% B; 3.2−3.32 min, 100−0% B and 3.32−6 min, 0% B. Solvent A consisted of 5 mM ammonium acetate in water (pH 7.4), and solvent B was HPLC-grade acetonitrile. Injection volume was 1 μL. The nitrogen nebulizing gas flow was set at 1.5 L/min and the drying gas flow at 15 mL/min. The interface voltage was 4500 V. The temperature of the block heater was maintained at 400 °C and that of the desolvation line at 250 °C.

Table 1 summarizes the results where (i) CHI 1 is the first CHI value, obtained with DMSO solutions in plates received from the central repository (the whole set was measured with UV−vis detection and used for the first modeling); (ii) CHI_{pred} stands for CHI average prediction and corresponds to the average prediction over the 10 best SVM models in the iterative procedure; (iii) CHI 2 is the second CHI value, obtained with fresh solutions prepared from powders (except for those marked with asterisks) and measured for the 24 outliers (with LC/UV); and (iv) MS indicates whether the presence of the theoretical structure was confirmed (Y) or invalidated (N) by mass spectrometry.

Outlier Analysis. The first 21 outliers from the list (see Table 1) were experimentally confirmed to be consequences of various experimental problems and artifacts, many of which escaped direct observation at the initial high-throughput measurement stage. The reassessment was extended to three additional compounds beyond this list of 21 outliers, in order to check the proposed outlier selection criteria.

Identified problems include chemical degradation, which could be identified for six compounds: one lactone (outlier 16), two anhydrides (outliers 5 and 10), and three esters (outlier 3, 8, and 12) were hydrolyzed and the resulting degradation was found in MS. Out of the 21 compounds, only six had an experimental comment indicating eventual measurement complications: three had precipitated in the buffer or in the DMSO stock solution, one had several peaks, one had a large peak, and one had a peak of low intensity. In total, 15 compounds had experimental problems where no measurement complications had been detected.

In order to discuss the results, different compounds have been regrouped into the following categories: hydrolyzed compounds, solutions containing several products, structure not confirmed by MS, no correspondence between the different CHI measurements, and no experimental problems.

Hydrolyzed Compounds: Outliers 3, 5, 8, 10, 12, and 16. In all these cases, the MS spectrum of the hydrolyzed molecule is found, proving the chemical degradation. Such reactions are generally considered as slow[20] at pH = 7.4. However, water impurities may be contained in the DMSO stock solution due to its hygroscopic nature, and thus reaction may occur before the compound is placed in the buffer solution. For outliers 8 and 12, it seems the degradation is fast enough to occur during the second measurement, and thus two peaks are found during the second measurement of CHI. In both cases, it can be assumed that the lowest value corresponds to the acid and the higher value to the drawn structure. In the case of outliers 5 and 10, powder was not available to remake a fresh solution. It seems CHI measurements correspond in both cases to the hydrolyzed compound. In the case of outlier 3, it can be assumed that the first measured value (CHI = 6.7) corresponds to the acid. In the case of the lactone (outlier 16), the compound is not observed and only the hydrolyzed molecule is detected by MS. It can thus be assumed that the CHI values correspond to it.

Solutions Containing Several Products: Outliers 4, 6, 7, 9, 11, 14, 15, and 20. The compounds are detected by MS but with contaminants, indicating a possible degradation or impurity. Outliers 4 and 11 both have benzyl bromides, which may be hydrolyzed[21] or degraded. In the case of outlier 11, the problem is likely related to low solubility of the compound, and hence an impurity is measured in LC/UV−vis with a more intense peak. In the case of outlier 6, the theoretical structure seems to correspond to the CHI value of 99.8. In the case of outlier 15, the expected compound is confirmed by LC/MS but has no chromophore to be detected in LC/UV−vis. Thus, the measured CHI value probably corresponds to an impurity or a counterion coming out at the void time.

Theoretical Structure Not Confirmed by MS: Outliers 1, 2, 17, and 19. The compounds are not present during the experiment. It is impossible to conclude what may have happened and what is actually measured during the LC/UV experiment with the given information. Possibly, the compound was not soluble or the given powder did not contain the indicated compound due to a human error. In the case of outlier 17, a substructure of the theoretical structure is found in MS. This could have been an input or synthesis error. In the case of outlier 2, the absence may be related to the low solubility of the compound (measured as 2 μM in pH 7.4 buffer).

No Correspondence between Different CHI Measurements: Outliers 13, 18, and 21. The compounds are identified by MS, but no matching of the CHI values can be found and no other compounds are detected. Possibly some wells in the given microplates may have contained a wrong solution in the first measurement or the compounds were degraded during storage and these reactions are not fast enough to be observed during the second measurement, when the stock solutions are redone. In the case of outliers 13 and 21, the predicted values are qualitatively in better accord with the second measurements. In the case of outlier 18, it is questionable whether the compound is not hydrolyzed or degraded.

No Experimental Problems: Outliers 22−24. The compounds are detected in the expected ranges of retention times by LC/MS and both CHI measurements match. It seems these molecules are not well predicted and the discrepancy may originate from the limits of the modeling. We note that the outliers 22 and 24 are above the highest calibration value (valerophenone, CHI = 96.4).

Extreme Values of CHI. CHI is derived from the ICHI, which corresponds to the percentage of acetonitrile needed to achieve an equal distribution between the two phases. It is calibrated on a set of compounds for which the ICHI is known and the ICHI is effectively bounded between 0 and 100.

dx.doi.org/10.1021/ac403544k | Anal. Chem. 2014, 86, 2510−2520

However, as the CHI is a retention time converted to an ICHI scale, it can have values outside the range 0−100.

Several outliers confirmed to have experimental problems have a negative value and it was observed that their CHI corresponds to the void time of the column; thus no actual measurement of the molecule's hydrophobicity is achieved. It can only be concluded that these have very low hydrophobicity. In the remaining molecules of the database, three such cases with values below 0 are found (structures are provided in Supporting Information section 4) and were thus discarded from the final modeling data set.

The 57 cases above 100 CHI units have been kept (excluding outlier 9), as these CHI value convey physicochemical meaningful differences between the compounds. Indeed, a retention time can be unambiguously measured: no metrological problem is expected. For this range of CHI, it can be assumed that a compound with a lower CHI than another has indeed a lower hydrophobicity. However, the assumption of a linear relationship between isocratic chromatographic hydrophobicity index and log D[8] is obviously wrong.

Outlier Dependence on the Modeling Protocol. The sensitivity of the outlier list with respect to the machine learning technique was assessed by ranking compounds according to the average errors reported by alternative PLS regression models obtained with Weka 3.7.6[16] and respective SQS[17] models. The PLS models were generated with varying number of components from 2 to 20 with a step of 2. SQS models were built on eight descriptor spaces known for their good predictive proficiency in SVM fitting. The 10 PLS models used were selected on the criteria of equivalent statistics to best model, low number of components, and different types of descriptors. The eliminate-and-refit approach was also used on PLS.

The other machine learning methods are also able to find most of these outliers, picked on the basis of SVM models. These were primarily run to cross-check whether outlier detection would be strongly impacted by the choice of machine learning protocols. This is not the case. The outlier lists obtained by use of PLS or SQS were largely consistent with the one obtained with SVM.

FINAL CONSENSUS MODEL

The compounds experimentally confirmed to have problems (21 compounds, see Table 1), compounds with CHI values below 0 (3 compounds), and all compounds with several peaks (36 compounds) were removed from the initial set. The "cleaned" data set of 485 compounds has been used to rebuild SVM models, re-exploring descriptor spaces and parameters. An external 5CV procedure was applied by splitting the initial set of molecules five times into five different folds. Best models were selected on the criterion of a 5CV RMSE better than a cutoff of 16. Only one model per descriptor space was kept. A y-randomization strategy[22] performed 20 times confirmed the significance of the selected models. In total, 81 models with 5CV-RMSE ranging from 14.5 to 16 are included in the consensus model (see Supporting Information section 7 for details).

It was observed that the best descriptor spaces were covering small fragments. The best descriptor space is an atom-centric fragmentation colored by atomic symbols with a range of 2−3 atoms and the use of formal charges. This might be related to the diversity of the molecules, which do not allow the extraction of more complex description, or to the additive character of hydrophobicity.[23]

An external test set of 195 molecules from the literature[7,8,24−26] was used to evaluate the generalization of the consensus model. Care was taken to have the most similar experimental conditions: (i) The pH varies from 7 to 7.4. (ii) A reversed-phase C18 column with a gradient of acetonitrile/buffered water was used in all cases. (iii) Calibration was slightly different in two cases;[7,26] hence, an equation was established to convert the values. (iv) Compounds were detected by UV−vis in most cases and by mass spectroscopy[25] for six molecules.

The model performs reasonably on the external test set with a RMSE of 16.4 and a determination coefficient R^2_{det} of 0.6 (see Supporting Information section 5 for details). It is not surprising to obtain worse results on the external test set than expected from cross-validation experiments. The main difference is that the former data set is issued from the literature whereas the latter is issued from the same laboratory. For data coming from literature, it is not possible to exclude some variation in the experimental setup, the least of it being that the calibration parameters of the CHI vary from one article to the other. The compounds measured by MS also notably differ from the other errors (see Supporting Information section 5 for details).

CONCLUSION

To conclude, we suggest the use of QSPR modeling to control the quality of HTS experiments. In this paper, we present the largest homogeneous data set of experimentally measured CHI values. We also propose an algorithm to list, on the basis of QSPR modeling, outliers that are likely to represent cases of severe and hidden experimental error. With this algorithm, we were able to pinpoint experimental problems for 21 compounds. These problems could not be detected during the experimental screening and they represented about 4% of the database. The final model was produced from reliable data and is publically available. The model was used to annotate the whole CN.

It is our belief that removal of outliers should not be done automatically (typical strategy in QSAR/QSPR) and outliers should bring chemists to reflect on their work. Their proper analysis demands a synergy between experimental screening teams and chemoinformatics modeling teams. The cost of a QSPR study is negligible compared to a screening campaign. The discrepancies observed between QSPR estimates and screening results are useful to detect experimental problems otherwise invisible. Such interplay could be a useful addition to regulatory tests such as those mentioned in REACH.

ASSOCIATED CONTENT

Supporting Information

Additional text, four tables, and two figures describing (1) calibration compounds and their associated ICHI values, (2) statistical parameter definitions, (3) parameters of 10 SVM models used for eliminate-and-refit protocol to detect outliers, (4) structures and CHI values of three compounds below 0 after removal of outliers, (5) experimental versus predicted value of CHI on external test set for the consensus model, (6) availability of the model for end users (http://infochim.u-strasbg.fr/webserv/VSEngine.html), and (7) descriptor spaces, ISIDA Fragmentor2012 options, libsvm options, and statistics of the 81 models used in the consensus model; and a listing of

dx.doi.org/10.1021/ac403544k | Anal. Chem. 2014, 86, 2510−2520

the CHI training set containing all measured values. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail varnek@unistra.fr.

Author Contributions

‡F.R. and P.G. contributed equally to the work

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank the CNRS PIR project for financial support and the Centre of High-Performance Computing, Informatics Department, University of Strasbourg (France), for computational facilities.

REFERENCES

(1) Yan, B.; Fang, L.; Irving, M.; Zhang, S.; Boldi, A. M.; Woolard, F.; Johnson, C. R.; Kshirsagar, T.; Figliozzi, G. M.; Krueger, C. A.; Collins, N. *J. Comb. Chem.* **2003**, *5* (5), 547−559.

(2) Lemoff, A.; Yan, B. *J. Comb. Chem.* **2008**, *10* (5), 746−751.

(3) Hansch, C.; Leo, A.; Mekapati, S. B.; Kurup, A. *Bioorg. Med. Chem.* **2004**, *12*, 3391−3400.

(4) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.

(5) Meyer, H. *Arch. Exp. Pathol. Pharmakol.* **1899**, *42*, 109−118.

(6) Hansch, C.; Fujita, T. *J. Am. Chem. Soc.* **1964**, *86* (8), 1616−1626.

(7) Valkò, K.; Bevan, C.; Reynolds, D. *Anal. Chem.* **1997**, *69* (11), 2022−2029.

(8) Valkò, K.; Du, C. M.; Bevan, C.; Reynolds, D. P.; Abraham, M. H. *Curr. Med. Chem.* **2001**, *8* (9), 1137−1146.

(9) Valkò, K.; Slégel, P. *J. Chromatogr.* **1993**, *631* (1-2), 49−61.

(10) Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D. *Mol. Inf.* **2010**, *29* (12), 855−868.

(11) Chang, C. C.; Lin, C.-J. *ACM Trans. Intell. Syst. Technol.* **2011**, *2* (3), 27:1−27:27.

(12) Barnett, V.; Lewis, T. *Outliers in Statistical Data*, 3rd ed.; John Wiley & Sons.: New York, 1994.

(13) Ahlers, J.; Stock, F.; Werschkun, B. *Environ. Sci. Pollut. Res.* **2008**, *15*, 565−572.

(14) ChemAxon JChem, Calculator plugin. http://www.chemaxon.com.

(15) Dietterich, T. G. *Neural Comput.* **1998**, *10* (7), 1895−1923.

(16) Hall, M.; Eibe, F.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. *ACM SIGKDD Explor. Newsletter* **2009**, *11* (1), 10−18.

(17) Horvath, D.; Bonachera, F.; Solov'ev, F.; Gaudin, C.; Varnek, A. *J. Chem. Inf. Model.* **2007**, *47* (3), 927−939.

(18) Weaver, S.; Gleeson, M. P. *J. Mol. Graphics Modell.* **2008**, *26* (8), 1315−1326.

(19) Baskin, I. I.; Kireeva, N.; Varnek, A. *Mol. Inf.* **2010**, *29* (8−9), 581−587.

(20) Clayden, J.; Greeves, N.; Warren, S.; Wothers, P., *Organic Chemistry*, 1st ed.; Oxford University Press: Oxford, U.K., 2001.

(21) Vitullo, V. P.; Sridharan, S.; Johnson, L. P. *J. Am. Chem. Soc.* **1979**, *101* (9), 2320−2322.

(22) Rücker, C.; Rücker, G.; Meringer, M. *J. Chem. Inf. Model.* **2007**, *47* (6), 2345−2357.

(23) Ghose, A. K.; Crippen, G. M. *J. Comput. Chem.* **1986**, *7* (4), 565−577.

(24) Plassa, M.; Valkò, K.; Abraham, M. H. *J. Chromatogr. A* **1998**, *803* (1−2), 51−60.

(25) Camurri, G.; Zaramella, A. *Anal. Chem.* **2001**, *73* (15), 3716−3722.

(26) Fuguet, E.; Ràfols, C.; Bosch, E.; Rosés, M. *J. Chromatogr. A* **2007**, *1173*, 110−119.