



QSPR modelling -a valuable support in HTS quality control

Fiorella Ruggiu, Patrick Gizzi, Jean-Luc Galzi, Marcel Hibert, Jacques Haiech, Igor Baskin, Dragos Horvath, Gilles Marcou, Alexandre Varnek

► To cite this version:

Fiorella Ruggiu, Patrick Gizzi, Jean-Luc Galzi, Marcel Hibert, Jacques Haiech, et al.. QSPR modelling -a valuable support in HTS quality control. *Analytical Chemistry*, 2014, 10.1021/ac403544k . hal-02196042

HAL Id: hal-02196042

<https://hal.science/hal-02196042>

Submitted on 26 Jul 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

QSPR modelling – a valuable support in HTS quality control

Fiorella Ruggiu^{1*}, Patrick Gizzi^{2*, 5}, Jean-Luc Galzi^{2, 5}, Marcel Hibert^{3, 5},
Jacques Haiech^{3, 5}, Igor Baskin^{1, 4}, Dragos Horvath¹, Gilles Marcou¹,
Alexandre Varnek^{1 *}

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

² Laboratoire de Biotechnologie et Signalisation Cellulaire (Plate-forme TechMed^{ILL}), UMR 7242 CNRS/Université de Strasbourg, Ecole Supérieure de Biotechnologie Strasbourg, 67412 Illkirch Graffenstaden, France

³ Laboratoire d'Innovation Thérapeutique, UMR7200 CNRS/Université de Strasbourg, Faculté de Pharmacie, 74 route du Rhin, 67401 Illkirch, France

⁴ Lomonosov Moscow State University, Moscow 119991, Russia

⁵ Laboratory of Excellence Medalis

* These authors contributed equally to the work

* Corresponding author, e-mail: varnek@unistra.fr

Keywords

HTS, QSPR, QSAR, Quality control, outliers, Hydrophobicity, Lipophilicity, CHI, Chromatographic Hydrophobicity Index

Abstract

The evaluation of important pharmacokinetic properties such as hydrophobicity using High Throughput Screening (HTS) methods is a major issue in drug discovery. In this article, we present the measurement of the Chromatographic Hydrophobicity Index (CHI) on a subset of the French chemical library, the “Chimiothèque Nationale” (CN). The data was used in QSPR modelling 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 reanalysed 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 the cooperation between experimental and theoretical teams to improve results. The corrected data was used to produce a model, which is freely available on our web server <http://infochim.u-strasbg.fr/webserv/VSEngine.html>

1 Introduction

Since the advent of robotized biological testing in the nineties, 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 since 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 only go as far as automated recording followed by error-prone machine interpretation of analysis results. Time and resources for in-depths structural analysis is lacking – therefore, standard purity measures are necessary, but hardly sufficient^{1, 2}. In standard Liquid Chromatography – Mass Spectroscopy (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) that 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 nation-wide 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, highlight and fix various pitfalls due to the peculiar nature of this collection. PIR was aimed at annotating the CN in 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 non-druglike molecules as well. The CNE molecules were then cherry-picked and submitted to standard 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 physico-chemical properties was carried out at the TechMed^{ILL} Platform in Strasbourg. Hydrophobicity – the first measured property, and the one concerned by this article – is an important property for medicinal chemists³. It is widely used as a criterion for acceptable drug solubility and permeability⁴. It has been shown to be related to ADME/T properties for over a century⁵. It has classically been evaluated by the octanol-water partition coefficient LogPo/w after the proposal of Hansch and Fujita⁶ and measured by the shake-flask method. However, this method is time-consuming and a modern HTS method using 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 LogPo/w measures⁹. However, this measure requires testing several mobile phases with different organic solvent percentages, thus being time and resources 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 ICHI⁷ and to logPo/w⁸. 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 the linear relation between the 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. QSPRs are mathematical models fitted on the data which 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 such as to ensure that, for each training compound (of known property Y), the model will return a predicted Ypred very close to Y (following the classical least-squares principle). The molecular descriptors used in this study are the ISIDA property-labelled fragment counts¹⁰. Fitting was performed using mainly 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 modelling. We define *outliers* as compounds for which their calculated property value Ypred 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 which is numerically distant from the rest of the data¹². We propose a method for their systematic annotation and then to submit them to in-depth experimental scrutiny. Since the observed discrepancies between Y and Ypred were much higher than the expected models imprecisions, and yet independent on modelling premises it was hypothesized that this could be due to real differences in molecular structures: the actual molecule returning the measured Y might not correspond to the nominal structure for which Ypred 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 (minimizing modelling artefacts 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 to potential experimental pitfalls, focussing the need for in-depth further analysis to the potentially “pathological” items. This could be an important first step towards the use of QSPR approaches for regulatory purposes, instead of experimental measurements, as envisaged by the REACH project¹³.

This paper is organised in order to follow the chronology of the different experimental and modelling steps within the study. First the experimental protocol and results of the CHI measurements is presented (see §2), followed by an outline of the computational procedures (§3) the outlier management section (see §4). Outlier management contains the initial building of the models, the modelling protocol for the identification of the outliers (§4.1), their experimental validation (§4.2) and a presentation of the results with a discussion (§4.3). Finally, the consensus model (see §5), build after removal of outliers and doubtful molecules from the set is presented, followed by a conclusion section.

2 CHI measurements

The 640 CNE compounds were received in 8 microplates containing 10 mM 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 μ m Luna C18(2) column (50 x 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 pH 7.4 ammonium acetate in water and solvent B was HPLC grade acetonitrile (Sigma-Aldrich CHROMASOLV). The detection wavelengths were 254 and 230 nm.

First, a solution with 10 reference compounds with known ICHI values (see Supporting Information 1.) was injected onto the HPLC to generate a calibration line from their retention times (see Figure 1). The concentration of the mixture was 0.2 mg/mL for each compound and the injected volume was 3 μ L. A typical chromatogram of the standard solution is represented in Figure 2. The test compounds were analysed on the same system. The 10 mM DMSO stock solutions were diluted to 200 μ M 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).

Insert Figures 1 and 2 here

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 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 molecule for unknown and probably undefined reasons (presumably compound insolubility or unstable in DMSO, 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.

Insert Figure 3 here

3 Computational procedure

The computational workflow used in this work is given on Figure 4. Steps 1-5 are describes in Section 3 whereas steps 6-8 are reported in Section 5.

Insert Figure 4 here

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 micro-species at pH=7.4 with ChemAxon's Calculator plugin¹⁴.

Descriptors Calculation. ISIDA property-labelled descriptors¹⁰, a type of fragment count descriptors, were calculated. Sequences, extended augmented atoms and triplets were computed on the molecular graph which has been "coloured" 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 package¹¹ was used for the generation of epsilon-SVM regression models with a linear kernel. Epsilon 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 techniques¹⁵ to avoid overfitting. The final model selection criterion therefore was the 5-fold cross-validated root-mean-squared error (5CV-RMSE) (See Supporting Information 2. for details on statistical parameters).

PLS and SQS regression models issued from selected pools of descriptors were also built for the comparison purposes.

Model Selection. Totally, $2772 \times 28 = 77616$ individual models (each corresponding to particular descriptor pool and a particular value of cost parameter) have been obtained for a given dataset. Several “best” models has been selected according to 5CV-RMSE. All selected models were used for consensus predictions on the external test set: for each molecule, CHI value was calculated as an arithmetic average of predictions made by selected individual models.

Outlier identification protocol. In this section we discuss identification of recurrent outliers observed in different modelling 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 modelling – was gathered using an *eliminate-and-refit* protocol on the basis of N best models. At each step of the prediction for a given data point is considered anomalous if its calculations 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: $C_{out} = 2 \times \max(5CV-RMSE)$. 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 modelling dataset and the N models are refitted. The operational parameters are not re-optimized.
3. The experimentally measured CHI value in discrepancy with the prediction is challenged, by a thorough re-analysis of the compound (see §4.2), with four possible outcomes:
 - a) the initial CHI value is proven wrong, and a correct estimate is found instead,
 - b) the initial CHI value is proven wrong, and the renewed attempt to measure the property fails,
 - c) the initial CHI value is reconfirmed, but structural analysis shows that the actual compound corresponding to the detected peak is not the nominal structure from the electronic database,

- d) both the initial CHI value and compound structure are reconfirmed – thus, this is a modelling problem.
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, b, c listed above).

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 domain¹⁶. 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 significantly skew the calculated values for other compounds. When eliminating one molecule 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 to converge towards problematic molecules identified by different points of views.

4 Outlier detection, validation and analysis

4.1 Outlier detection

10 models out of 77616 built on the parent set of 545 compounds were selected according to 5CV-RMSE. The best of them involves atom-centric fragments coloured by atomic symbols with a range of 2 to 4 atoms and 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).

Insert Figure 5 here

The CNE set is the biggest collection of CHI values found in 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 modelled due to multiple protocol incoherencies was discarded. A closer analysis of the structure of those molecules showed that certain 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 relatively poor model performance is due to including into training set the molecules for which some experimental problems were detected (blue portion of the pie in Figure 3), the modelling was performed on the set of 418 molecules measured without any complications (green portion of the pie in Figure 3). We didn't observe any significant improvement of performance, 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 was 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 in §3 has been applied for 10 best SVM models (see the models parameters in Supporting Information 3). This lead to detection 24 outliers listed in Table 1. Unsurprisingly, outliers detected at fitting stage also behave erratically during 5CV (see Figure).

Insert Table 1 here

To ensure the outliers did not contain unique features which would make them fundamentally different from the others in the training, 1-SVM¹⁷ using a linear kernel was applied at varying ν parameter. The outlier distribution is homogeneous within the dataset. The percentage coverage within the outliers corresponds to the percentage coverage within the dataset. 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.

4.2 Experimental reassessment of outliers

The experimental check of compounds annotated as outliers were done by the TechMed^{ILL} Platform. CHI of the compounds identified as outliers were measured a second time (CHI 2 in Table 1) and solutions were submitted to mass spectrometry re-characterisation in order to explain differences found between experimental and predicted CHI values. Fresh DMSO stock solutions were prepared from powders except for 4 compounds for which powder was not available (indicated by a * in Table 1). The powder should contain less impurities and eventual chemical degradation are less likely to occur than in the stock solution.

Firstly, these solutions were used to determine the CHI values again by the same procedure explained previously (see **Erreur ! Source du renvoi introuvable.**). It permits to check whether the stock solutions distributed by the CN had problems. Secondly, a LC-MS characterisation 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 to control if 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, thus, 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 x 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; 3-3.30 min, 100% B; 3.2-3.32 min, 100-0% B and 3.32-6 min, 0% B. Solvent A consisted of 5 mM pH 7.4 ammonium acetate in water 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. 4500 V were used for the interface voltage. The temperature of the block heater was maintained at 400 °C and the one of the desolvation line at 250 °C.

Table 1 summarizes the results where:

- 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 a *) and measured for the 24 outliers (with LC-UV);
- MS indicates whether the presence of the theoretical structure was confirmed by mass spectrometry (indicated by Y) or invalidated (indicated by N)

4.3 Outlier analysis

The first 21 outliers from the list (see Table 1) were experimentally confirmed to be consequences of various experimental problems and artefacts, 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 6 compounds: one lactone (outlier 16), two anhydrides (outlier 5 and 10) and three esters (outlier 3, 8 and 12) were hydrolysed and the resulting degradation was found in MS. Out of the 21 compounds, only 6 had an experimental comment indicating eventual measurement complications: 3 had precipitated in the buffer or in the DMSO stock solution, 1 had several

peaks, 1 had a large peak and 1 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 6 categories: Hydrolysed compound, solutions containing several products, structure not confirmed by MS, no correspondence between the different CHI measurements and no experimental problems.

Hydrolysed compounds: *Outlier 3, 5, 8, 10, 12, 16.* In all these cases, the MS spectrum of the hydrolysed molecule is found, proving the chemical degradation. Such reactions are generally considered as slow¹⁸ 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 placing the compound in the buffer solution. For outlier 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 outlier 5 and 10, powder was not available to remake a fresh solution. It seems CHI measurements correspond in both cases to the hydrolysed 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 hydrolysed molecule is detected by MS. It can thus be assumed that the CHI values correspond to it.

Solutions containing several products: *Outlier 4, 6, 7, 9, 11, 14, 15, 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 hydrolysed¹⁹ or degraded. In the case of outlier 11, the problem is likely related to a 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: *Outlier 1, 2, 17, 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 μ M in pH 7.4 buffer).

No correspondence between different CHI measurements: *Outlier 13, 18, 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 the storage and these reactions are not fast enough to be observed during the second measurement, when redoing the stock solutions. In the case of outlier 13 and 21, the predicted values are

qualitatively in better accordance to the second measurements. In the case of outlier 18, it is questionable whether the compound is not hydrolysed or degraded.

No experimental problems: *Outlier 22, 23, 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 origin from the limits of the modelling. 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. However, as the CHI is a retention time converted to an ICHI scale, it can have values outside of the range 0-100.

Several outliers confirmed to have experimental problems have a negative value and it was observed that their CHI correspond to the void time of the column, thus, no actual measurement of the molecule's hydrophobicity is done. It can only be concluded these have a 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 4.) and were thus discarded from the final modelling dataset.

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 to the isocratic chromatographic hydrophobicity index and to LogD⁸ is obviously wrong.

Outlier dependence on the modelling 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 Partial Least Square (PLS) regression models obtained with Weka 3.7.6²⁰ and respectively Stochastic QSAR Sampler (SQS)²¹ 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 8 descriptor spaces known for their good predictive proficiency in SVM fitting. 10 PLS models used were selected on the criteria of equivalent statistics to best model, low number of components and different type 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 using PLS or SQS were largely consistent with the one obtained with SVM.

5 Final consensus model

The compounds experimentally confirmed to have problems (21 cmpds, see Table 1), compounds with CHI values below 0 (3cmpds) and all compounds with several peaks (36 cmpds) were removed from the initial set. The “cleaned” dataset 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 5 times into 5 different folds. Best models were selected on the criteria of a 5CV RMSE better than a cut-off of 16. Only one model per descriptor space was kept. A *y*-randomisation strategy²² 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 7. for details).

It was observed that the best descriptor spaces were covering small fragments. The best descriptor space is an atom-centric fragmentation coloured by atomic symbols with a range of 2 to 3 atoms and with 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 hydrophobicity²³.

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:

- The pH varies from 7 to 7.4.
- A reversed-phase C18 column with a gradient of acetonitrile/buffered water was used in all cases.
- Calibration was slightly different in two cases^{7, 26}, hence, an equation was established to convert the values.
- Compounds were detected by UV-Vis in most cases, and by mass spectroscopy²⁵ for 6 molecules.

The model reasonably performs on the external test set with a RMSE of 16.4 and a determination coefficient R^2_{det} of 0.6 (see Supporting Information 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 dataset 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 5. for details).

This model was used to annotate the CN and is freely available online: <http://infochim.u-strasbg.fr/webserv/VSEngine.html> (see Supporting Information 6. for instructions).

6 Conclusion

To conclude, we suggest the use of QSPR modelling to control the quality of HTS experiments. In this article, we present the largest homogeneous dataset of experimentally measured CHI values. We also propose an algorithm to list, based on QSPR modelling, 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 using 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 the chemists to reflect on their work. Their proper analysis demands a synergy between experimental screening teams and chemoinformatics modelling teams. The cost of a QSPR study is negligible compared to a screening campaign. The discrepancies observed between the QSPR estimates and the 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.

7 Acknowledgments

We thank the CNRS PIR project for financial support and the centre of High-Performance computing of the Informatics department of the University of Strasbourg (France) for the computational facilities.

8 Abbreviations used

ADME/T, Absorption Distribution Metabolism Excretion / Toxicity; CHI, Chromatographic Hydrophobicity Index; CN, Chimiothèque Nationale (French chemical library); CNE, Chimiothèque Nationale Essentielle (a subset of the CN); 5CV, 5-fold Cross-Validation; HPLC, High Pressure Liquid Chromatography; HTS, High-Throughput Screening; ICHI, Isocratic Chromatographic Hydrophobicity Index; LC, Liquid Chromatography; MS, mass spectroscopy; PIR, Projet Interdisciplinaire de Recherche (Interdisciplinary Research Project); PLS, Partial Least Square regression; QC, Quality Control; QSAR, Quantitative Structure-Activity Relationships; QSPR, Quantitative Structure-Property Relationships; RMSE, Root-Mean Squared Error; REACH, Registration, Evaluation, Authorisation and Restriction of Chemicals; SQS, Stochastic QSAR Sample; SVM, Support Vector Machine; UV-Vis, Ultra-Violet and Visible spectroscopy;

9 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., Quality Control in Combinatorial Chemistry: Determination of the Quantity, Purity, and Quantitative Purity of Compounds in Combinatorial Libraries. *J. Comb. Chem.* **2003**, *5*, 547-559.

2. Lemoff, A.; Yan, B., Dual Detection Approach to a More Accurate Measure of Relative Purity in High-Throughput Characterization of Compound Collections. *J. Comb. Chem.* **2008**, *10*, 746-751.
3. Hansch, C.; Leo, A.; Mekapati, S. B.; Kurup, A., QSAR and ADME. *Bioorg. Med. Chem.* **2004**, *12*, 3391-3400.
4. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **2001**, *46*, 3-26.
5. Meyer, H., Zur Theorie der Alkoholnarkose. Erste Mittheilung. Welche Eigenschaft der Anästhetica bedingt ihrer narkotische Wirkung? *Archiv f. experiment. Pathol. u. Pharmacol.* **1899**, *42*, 109-118.
6. Hansch, C.; Fujita, T., ρ - σ - π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86* (8), 1616-1626.
7. Valkò, K.; Bevan, C.; Reynolds, D., Chromatographic Hydrophobicity Index by Fast-Gradient RP-HPLC: A high-Throughput alternative to log P/log D. *Anal. Chem.* **1997**, *69*, 2022-2029.
8. Valkò, K.; Du, C. M.; Bevan, C.; Reynolds, D. P.; Abraham, M. H., Rapid method for the estimation of octanol/water partition coefficient (log Poct) from gradient RP-HPLC retention and a hydrogen bond acidity term (Sigma-alphaH2). *Curr. Med. Chem.* **2001**, *8* (9), 1137-1146.
9. Valkò, K.; Slégel, P., New chromatographic hydrophobicity index (ϕ_0) based on the slope and intercept of the log k' versus organic phase concentration plot. *J. Chromatogr.* **1993**, *631*, 49-61.
10. Ruggiu, F.; Marcou, G.; Varnek, A.; Horvath, D., ISIDA Property-Labelled Fragment Descriptors. *Mol. Inf.* **2010**, *29*, 855-868.
11. Chang, C. C.; Lin, C.-J., LIBSVM : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2011**, *2* (3), 27:1-27:27.
12. Barnett, V.; Lewis, T., *Outliers in Statistical Data*. 3rd edition ed.; John Wiley & Sons.: 1994.
13. Ahlers, J.; Stock, F.; Werschkun, B., Integrated testing and intelligent assessment—new challenges under REACH. *Environ Sci Pollut Res* **2008**, *15*, 565-572.
14. ChemAxon JChem - Calculator plugin. <http://www.chemaxon.com>.
15. Dietterich, T. G., Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **1998**, *7*, 1895-1923.
16. Weaver, S.; Gleeson, M. P., The importance of the domain of applicability in QSAR modeling. *J. Mol. Graphics Modell.* **2008**, *26* (8), 1315-1326.
17. Baskin, I. I.; Kireeva, N.; Varnek, A., The One-Class Classification Approach to Data Description and to Models Applicability Domain. *Molecular Informatics* **2010**, *29* (8-9), 581-587.
18. Clayden, J.; Greeves, N.; Warren, S.; Wothers, P., *Organic Chemistry*. 1st edition ed.; Oxford University Press: 2001.
19. Vitullo, V. P.; Sridharan, S.; Johnson, L. P., Neighboring ortho-carboxyl group participation and alpha-deuterium isotope effects in the hydrolysis of benzyl bromides. *J. Am. Chem. Soc.* **1979**, *101* (9), 2320-2322.
20. Hall, M.; Eibe, F.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H., The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **2009**, *11* (1), 10-18.
21. Horvath, D.; Bonachera, F.; Solov'ev, F.; Gaudin, C.; Varnek, A., Stochastic versus stepwise strategies for quantitative structure-activity relationship generation--how much effort may the mining for successful QSAR models take? *J Chem Inf Model* **2007**, *3*, 927-939.

22. Rücker, C.; Rücker, G.; Meringer, M., y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model* **2007**, 47 (6), 2345–2357.
23. Ghose, A. K.; Crippen, G. M., Atomic Physicochemical Parameters for Three-Dimensional Structure - Directed Quantitative Structure- Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. *J. Comput. Chem.* **1986**, 7 (4), 565-577.
24. Plassa, M.; Valkò, K.; Abraham, M. H., Determination of solute descriptors of tripeptide derivatives based on high-throughput gradient high-performance liquid chromatography retention data. *J. Chromatogr. A* **1998**, 803 (1-2), 51-60.
25. Camurri, G.; Zaramella, A., High-Throughput Liquid Chromatography/Mass Spectrometry Method for the Determination of the Chromatographic Hydrophobicity Index. *Anal. Chem.* **2001**, 73, 3716-3722.
26. Fuguet, E.; Ràfols, C.; Bosch, E.; Rosés, M., Determination of the chromatographic hydrophobicity index for ionisable solutes. *J. Chromatogr. A* **2007**, 1173 110-119.

10 **Figures captions**

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

Figure 2. Typical chromatogram of the standard solution

Figure 3. Experimental problems during CHI measurements on 640 molecules: in green, no problems, in red problems where no value could be determined and in blue problems but with a value

Figure 4. Computational workflow used in this work

Figure 5. Experimental vs. predicted CHI assessed at the fitting stage (a) and in 5-fold cross-validation (b) for the best SVM model (see section 4). The numbers indicate the outliers detected in the eliminate-and-refit protocol and listed in Table 1..... 20

Figures.

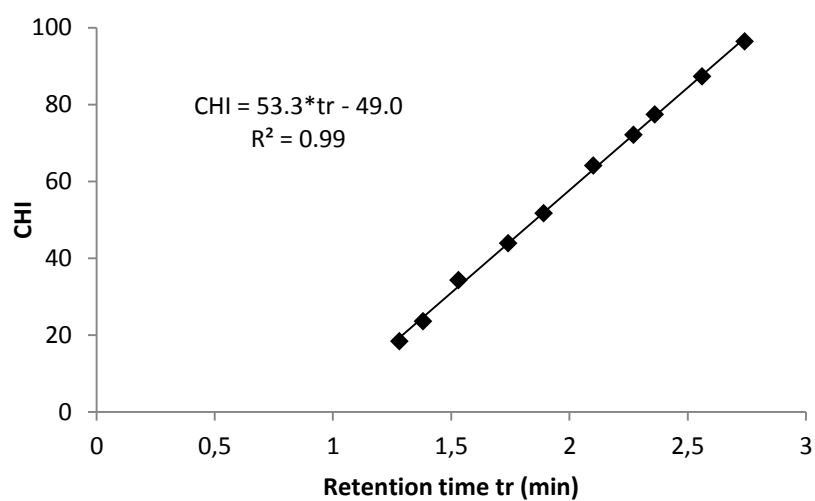


Figure 1. Calibration of the HPLC column: Relationship between retention times and CHI values

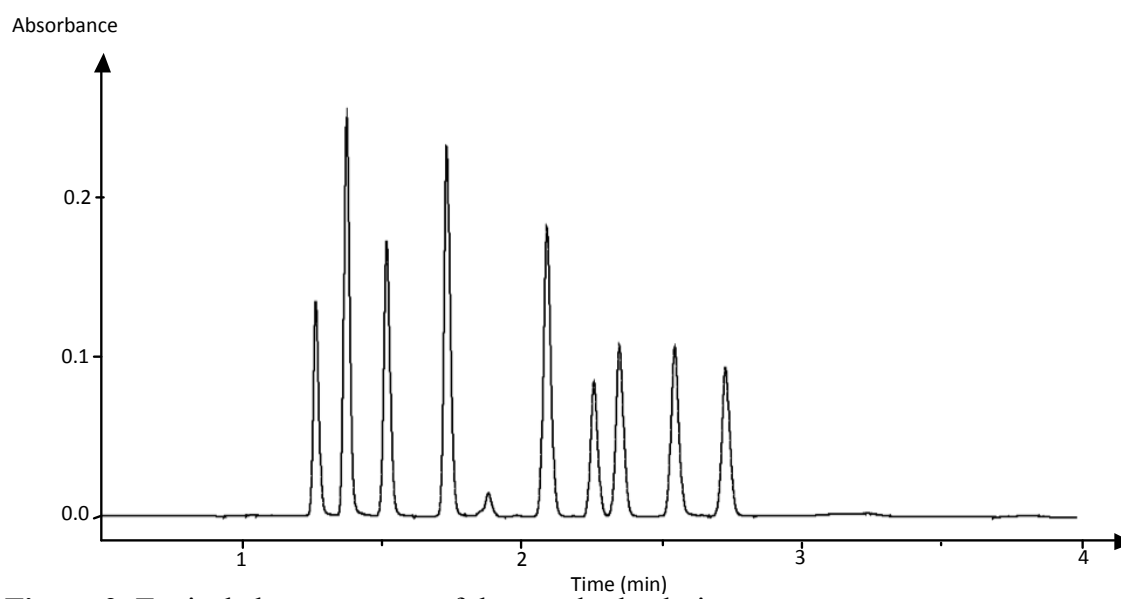


Figure 2. Typical chromatogram of the standard solution

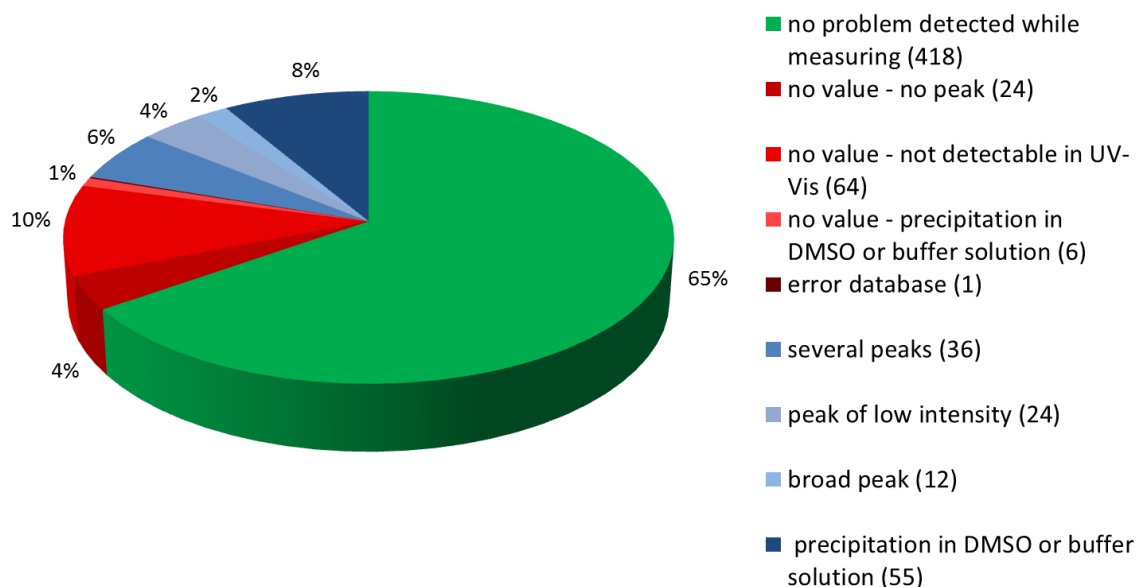


Figure 3. Experimental status of CHI measurements on 640 molecules: in green, no problems detected, in red failures to determine the CHI value and in blue measurements accompanied by observed side phenomena that may signal artefacts, all while nevertheless allowing some CHI value to be recorded.

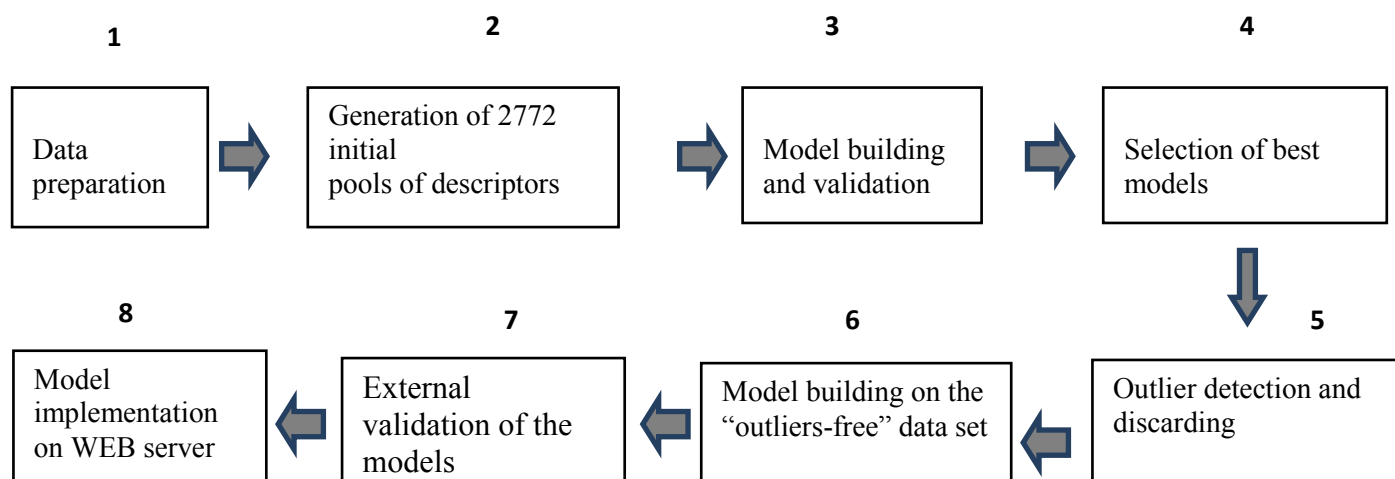


Figure 4. Computational workflow used in this work.

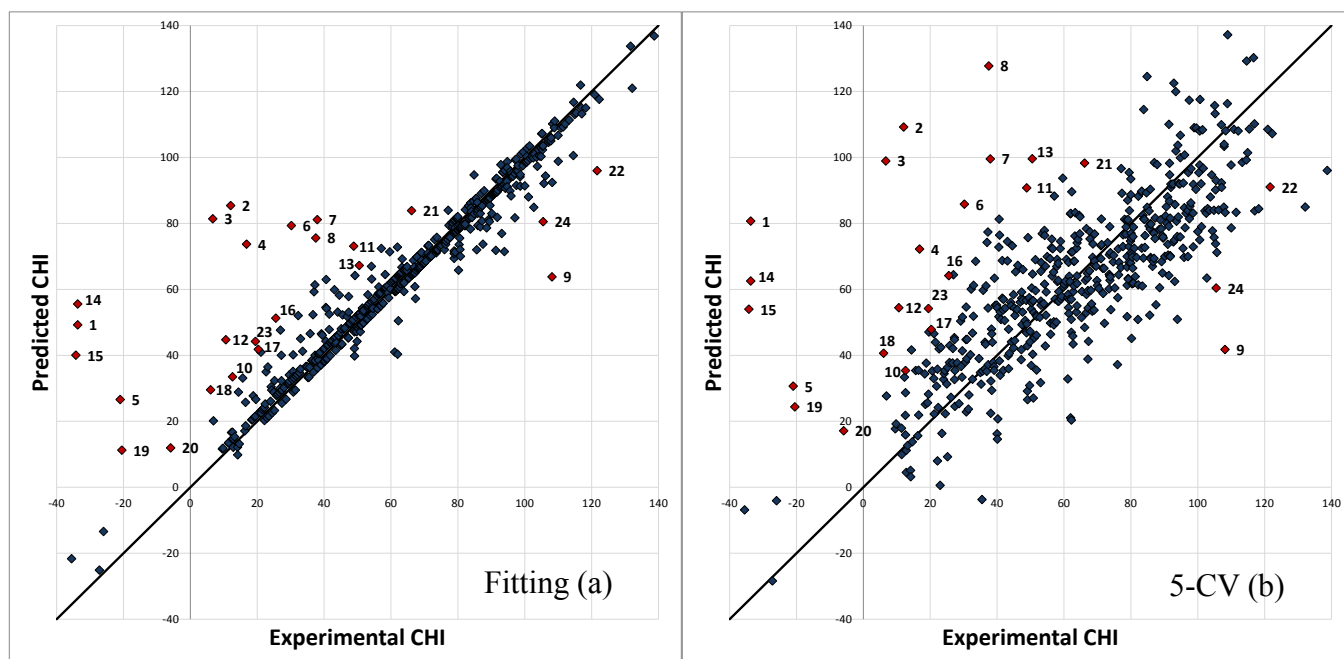
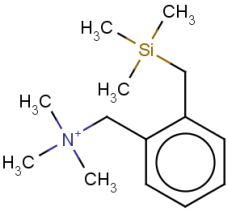
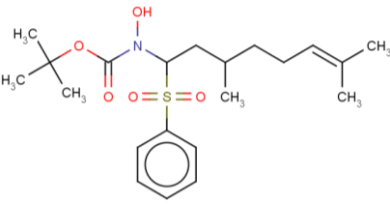
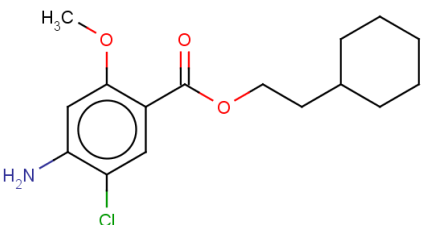
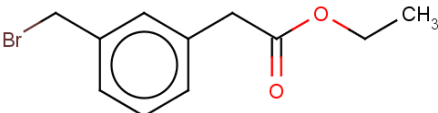
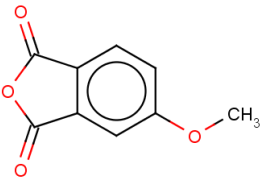
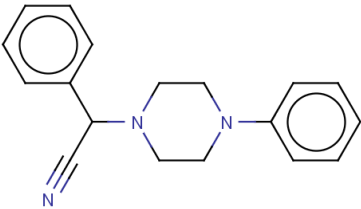
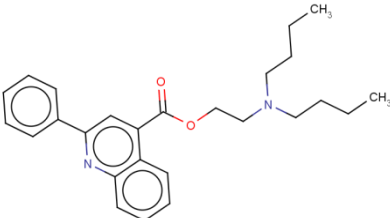
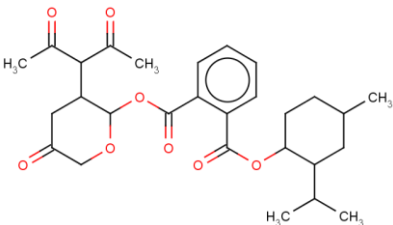
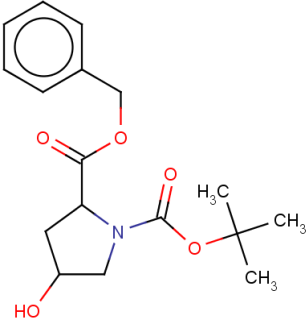
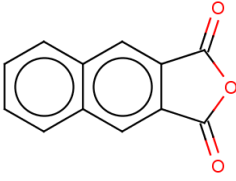


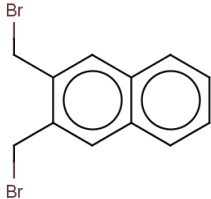
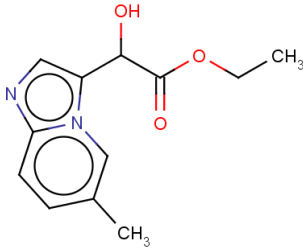
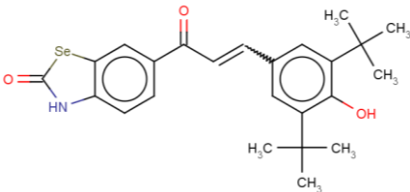
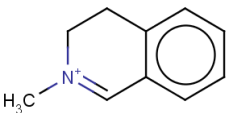
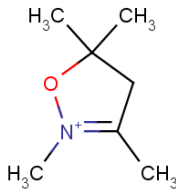
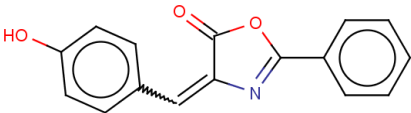
Figure 5. Experimental vs. predicted CHI assessed at the fitting stage (a) and in 5-fold cross-validation (b) for the best SVM model (see section 4). The numbers indicate the outliers detected in the *eliminate-and-refit* protocol and listed in Table 1.

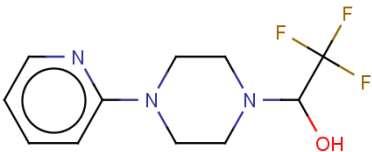
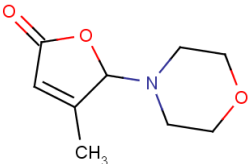
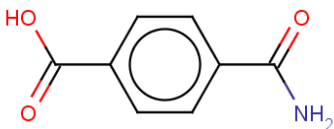
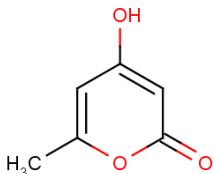
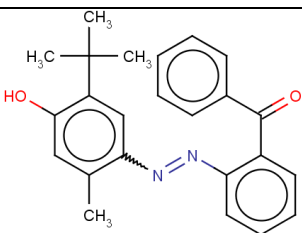
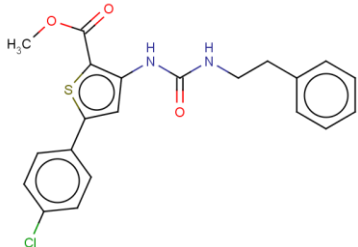
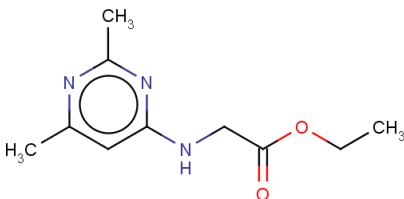
11 Tables

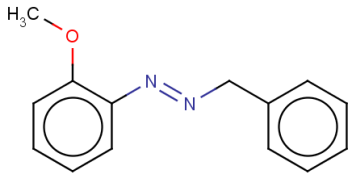
Table 1. Outliers list and experimental results

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
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	25.6	59.1	24.5	N

		detected by MS. The low value obtained for CHI1 experiment is explained by this hydrolysis.				
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
----	---	----------------------	-------	------	-------	---