

# Kernel-Based, Partial Least Squares Quantitative Structure-Retention Relationship Model for UPLC Retention Time Prediction: A Useful Tool for Metabolite Identification

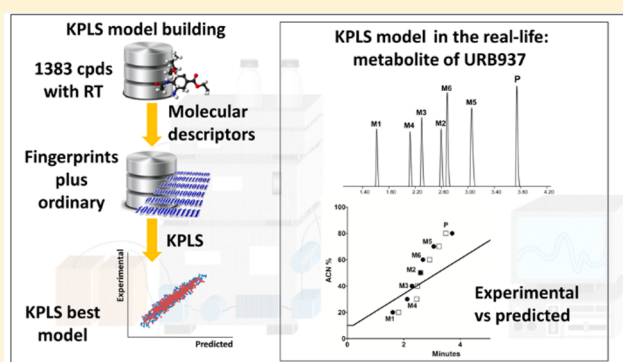
Federico Falchi,<sup>†,§</sup> Sine Mandrup Bertozzi,<sup>†,§</sup> Giuliana Ottonello,<sup>†</sup> Gian Filippo Ruda,<sup>†</sup> Giampiero Colombano,<sup>†</sup> Claudio Fiorelli,<sup>†</sup> Cataldo Martucci,<sup>†</sup> Rosalia Bertorelli,<sup>†</sup> Rita Scarpelli,<sup>†</sup> Andrea Cavalli,<sup>†,‡</sup> Tiziano Bandiera,<sup>†</sup> and Andrea Armirotti<sup>\*,†</sup>

<sup>†</sup>Drug Discovery and Development Department, Fondazione Istituto Italiano di Tecnologia, Via Morego 30, 16163 Genova, Italy

<sup>‡</sup>Department of Pharmacy and Biotechnology, University of Bologna, Via Belmeloro 6, 40126 Bologna, Italy

## S Supporting Information

**ABSTRACT:** We propose a new QSRR model based on a Kernel-based partial least-squares method for predicting UPLC retention times in reversed phase mode. The model was built using a combination of classical (physicochemical and topological) and nonclassical (fingerprints) molecular descriptors of 1383 compounds, encompassing different chemical classes and structures and their accurately measured retention time values. Following a random splitting of the data set into a training and a test set, we tested the ability of the model to predict the retention time of all the compounds. The best predicted/experimental  $R^2$  value was higher than 0.86, while the best  $Q^2$  value we observed was close to 0.84. A comparison of our model with traditional and simpler MLR and PLS regression models shows that KPLS better performs in term of correlation ( $R^2$ ), prediction ( $Q^2$ ), and support to MetID peak assignment. The KPLS model succeeded in two real-life MetID tasks by correctly predicting elution order of Phase I metabolites, including isomeric monohydroxylated compounds. We also show in this paper that the model's predictive power can be extended to different gradient profiles, by simple mathematical extrapolation using a known equation, thus offering very broad flexibility. Moreover, the current study includes a deep investigation of different types of chemical descriptors used to build the structure-retention relationship.



In 2004, Waters Inc. achieved one of the biggest advances in chromatographic science by introducing the first ultra performance liquid chromatography system (UPLC or UHPLC). For the first time, a commercially available instrument allowed the efficient use of stationary phases with a particle size of less than 2  $\mu\text{m}$ . Thanks to its ability to deal with very high backpressures (up to 1000 bar), this technology greatly increased the overall efficiency and productivity of LC laboratories, enabling much shorter separation times, higher peak capacity values, a significant reduction in solvent consumption, and most importantly a very high reproducibility of the retention time (RT) of a given compound.<sup>1,2</sup> This reproducibility can be achieved at any level,<sup>3</sup> from interday operations to batch-to-batch column performance evaluation, and it has become an important parameter to be considered when identifying unknown compounds. Several predictive models for RT are available for standard HPLC conditions<sup>4–6</sup> and some of them have become commercially available to aid method development. All these tools work using two different approaches. In tools like Chromsword<sup>7</sup> and ACD/ChromGenius,<sup>6</sup> the prediction is based on the molecular structure of the

compound itself and the method development is aided by an internal database of similar compounds of known RT. Other tools, such as Drylab,<sup>8</sup> assist method development starting from a small subset of preliminary experiments performed at different temperature, pH values, and gradient conditions. Method optimization is then guided by accurate RT prediction down to minor, but important, details, such as the expected peak width. Another field where the current topic is becoming more and more popular is proteomics, where peptide (thus protein) identification benefits from accurate RT prediction of peptides: several tools and models have been published<sup>9–11</sup> for this purpose. More in general, the metabolomics field would greatly benefit from a reliable RT prediction, as recently demonstrated by Creek et al.<sup>12</sup> with a quantitative structure-retention relationship (QSRR) method for hydrophilic interaction chromatography (HILIC) RT prediction. However, at present, very few studies report on UPLC/UHPLC RT

Received: May 27, 2016

Accepted: September 1, 2016

Published: September 1, 2016

prediction of small synthetic molecules showing the most diverse chemical heterogeneity. Several studies in this field are class-specific<sup>13–15</sup> and/or limited to a reduced data set,<sup>16</sup> although some are based on nonlinear tools such as artificial neural networks and support vector machine. Moreover, all these studies are based on calculated chemical properties such as theoretical AlogP or pK<sub>a</sub> and, therefore, predictive power issues are likely to arise,<sup>4</sup> as these descriptors are in turn calculated from other predictive models. Descriptors like fingerprints only depend on the molecular structure itself and thus do not suffer from this problem. An interesting approach to the problem is represented by the recent work of Thomaidis's group.<sup>17</sup> In this work, the authors used a large data set of compounds described with 22 different types of descriptors. With these descriptors the authors built linear (multiple linear regression) and nonlinear (artificial neural networks and support vector machine) robust models, paying attention to the selection of the descriptors, the choice of the training and test set, the external validation, and the outlier identification. In the present work we used the RT values of a large chemical library of 1383 synthetic compounds, comprising the most diverse chemical classes, to derive a fast kernel-based partial least squares (KPLS) model<sup>18</sup> using large collections of classical physicochemical and topological descriptors in combination with Canvas2D fingerprints for RT prediction. We then challenged our model with two real-life metabolite identification experiments, carefully evaluating the goodness of the model in predicting the retention time, the elution order of isomeric species and the predicted peak separation. Additionally, we showed how our KPLS model can be efficiently extended to different gradient conditions by simple extrapolation.

## ■ EXPERIMENTAL SECTION

**UPLC Data Collection.** All commercially available reagents and solvents were used as purchased from vendors without further purification. UPLC/MS analyses were conducted on a Waters Acquity UPLC/MS system (Waters Inc. Milford, MA) consisting of a single quadrupole detector (SQD) mass spectrometer equipped with an electrospray ionization interface and a photodiode array detector (PDA). The scan range was set to 110–650 *m/z* for both polarities (ESI+ and ESI–). The PDA range was 210–400 nm, and the UV purity and retention time were determined at the specific wavelength of 215 nm. The analyses were performed on an Acquity UPLC BEH C<sub>18</sub> column (100 mm × 2.1 mm i.d., particle size 1.7 μm) with a VanGuard BEH C<sub>18</sub> precolumn (50 mm × 2.1 mm i.d., particle size 1.7 μm) using 10 mM NH<sub>4</sub>OAc in H<sub>2</sub>O at pH 5 adjusted with acetic acid (AA) and 10 mM NH<sub>4</sub>OAc in acetonitrile (ACN)–H<sub>2</sub>O (95:5) at pH 5 (B) as mobile phase. After an initial hold for 0.2 min at 10% B, a linear gradient was applied to 90% B in 6 min, then from 90 to 100% B in 0.1 min, followed by a hold at 100% for 0.4 min. A 10 mM stock solution in dry DMSO was prepared for each test compound, and further diluted 20× in ACN–H<sub>2</sub>O (1:1) prior to analysis.

**Computational Chemistry. Structure Design and Optimization.** All computational chemistry studies were performed using Schrödinger software, release 2015-4, purchased from LLC, New York. All molecules were built using the Maestro graphical interface (Maestro, version 10.4) and subjected to the LigPrep module (LigPrep, version 3.6). LigPrep was set to produce a single low-energy 3D structure (Force Field OPLS 2005)<sup>19</sup> and multiple protonation/tautomeric states at pH 5.0

with Epik (Epik, version 3.4). In the case of chiral compounds with unknown stereochemistry, only one stereoisomer was generated. Moreover, only the most probable tautomer/protonation state was retained for each compound.

**Fingerprints and Ordinary Molecular Descriptors Calculation.** All structures were imported in Canvas (Canvas, version 2.6) and the MOLPRINT 2D, Radial, Dendritic, and Linear hashed fingerprints were generated with the default settings except for the use of a precision of 64 bit to avoid any collisions. A very detailed description of the Canvas fingerprints is provided by the work of Woody Sherman.<sup>20</sup> For all molecules, all the ordinary molecular descriptors available in Canvas (Physiochemical, Topological, LigFilter, and QikProp, from Small-Molecule Drug Discovery Suite 2015-4 and QikProp, version 4.6) were calculated. A complete list of all molecular descriptors was reported in the [Supporting Information](#). The set of molecular properties or descriptors obtained from various sources can often have a high degree of correlation. For model building, it is useful to remove the linear dependence in the properties, to speed up the calculation and reduce the numerical error. This can be achieved by selecting a smaller set of properties that adequately represents the larger set, in some measurable way. By using the “Selection dialog box” feature the properties whose values are identical in more than 90% of the molecules and the molecular descriptors with correlation above 0.9 were removed. Further details on Canvas can be found at <https://www.schrodinger.com/>.

**Multiple Linear Regression Analysis.** The classical multiple linear regression (MLR) models were built with Canvas using the above-mentioned ordinary descriptors. The data set was randomly separated into a training (70%) and an internal test set (30%). Each model was built by choosing the best subsets with a number of X variables from 1 to 10. We selected the option “weight models by R<sup>2</sup>” and default settings were used for the simulated annealing.

**Partial Least Squares Regression Analysis.** The classical partial least squares regression (PLS) model was built with Canvas using the ordinary descriptors. The data set was randomly separated into a training (70%) and an internal test set (30%). A “maximum number of PLS factors” of 5 and the “Autoscale X variables” option were selected.

**Kernel-Based PLS Model Building.** For each fingerprint and for the molecular properties, a principal components analysis (PCA) was performed using the PCA task available in Canvas. All models were created with the KPLS regression module available in Canvas. The maximum number of KPLS factor was set to 5, but this factor was automatically reduced if the standard deviation (SD) dropped below 0.4. The kernel nonlinearity was changed in the range 0.01–0.1. The retention time was set as the Y variable, and the fingerprints and/or the molecular properties were set as the X variable. The data set was randomly separated into a training (70%) and an internal test set (30%). For each model, the uncertainty was calculated using 10 bootstrapping cycles. Bootstrapping is performed by randomly sampling the training set with replacement to generate a new test set of the same size (that may include duplicates) and building a new model to predict the test set. The procedure can be repeated a user-specified number of times. The standard deviation from the original test set is then calculated and defined as the uncertainty.

**Synthetic Chemistry. Synthesis of Reference Metabolites M1–M6.** Detailed synthetic procedures for the preparation of

M1–M4 and M6 are reported in the [Supporting Information](#). MS was prepared as previously described.<sup>21</sup>

## RESULTS AND DISCUSSION

**Building of the Training Set.** Since 2010, our department produced a large library of compounds (>2500) belonging to different chemical classes. After their synthesis, all the compounds underwent extensive UPLC-UV-MS quality control analysis using reversed phase chromatography. All the observed RT values were duly recorded in an internal database. Thanks to this procedure's extremely high degree of standardization, we believed that this data set offered a valuable starting point for creating a KPLS statistical model to predict the UPLC retention time of any generic small molecule compound. We first evaluated the real reproducibility of the LC separation on a set of randomly chosen compounds, whose QCs were performed on different batches over a time span of 4 years (2011–2014, [Table S-1](#)). As reported, the RT values are extremely reproducible over time, with a relative standard deviation (RSD %) ranging from 2 to 5%. They thus offer a solid and robust experimental background for this kind of computational chemistry study. We then selected all the chemical compounds analyzed with the generic LC method described in the [Experimental Section](#) (1383 structures, *data set*). Using their corresponding RT values, we set up a KPLS model<sup>18</sup> on the basis of the ordinary and fingerprints descriptors. For each molecule in the data set, all the classical molecular descriptors (Physiochemical, Topological, LigFilter, and QikProp) and fingerprints (Linear, Radial, Dendritic, and MOLPRINT2D) available in the Canvas software were calculated. The data set was randomly separated into a training set (~2/3 of the compounds) and a test set (~1/3 of the compounds). Two different approaches are currently available to select the training/test set: (i) the random splitting and (ii) the “intelligent selection” as in the case of the Kohonen Self-Organizing Map Method, the Kennard-Stone method, the D-optimal design, the D-optimal onion design, sphere exclusion based methods, and K-Nearest neighbors (kNN). The random selection based methods are easily implemented but the selection is not rationally performed. On the contrary, the intelligent methods are time-consuming but they can increase the model predictive power. However, as reported by Tropsha et al.,<sup>22</sup> differences between the two approaches in terms of predictive power are minimal and random methods are well accepted. Therefore, random splitting methods were chosen in this study. Moreover, as shown in [Figure S-1](#), compounds belonging to both the training and test set are uniformly dispersed in the descriptor space and in the RT distribution profile. Furthermore, as reported in the data analysis section in the [Supporting Information](#) for each model, the estimated residual VS retention time (RT) plot did not show any heteroscedasticity. Finally, to test a possible influence of the training/test set selection on the final  $R^2$ , several changes in the seed number used for the random number generation were tested without a significant change in the final  $R^2$  of each generated model. The 70/30 ratio for the training/test set allowed the best predictive power on the external test set. A similarity/distance matrix was built (see [Figure S-2](#)) on the basis of the Tanimoto similarity index calculated with the linear fingerprints. As shown in [Figure S-2](#), the majority of the area in the plot is brown to orange colored, meaning that compound similarities fall below 0.1. Moreover, as the plot was built after having sorted compounds for increasing RTs (compound i + 1

has a higher RT than compound i) and blue areas are absent near the diagonal line, even compounds with similar RTs have a low structural similarity. Before building the models, a PCA<sup>23</sup> was carried out with traditional descriptors and fingerprints in order to collect the variance of each component and the cumulative variance. According to the PCA analysis, ordinary descriptors seem to explain more variance (PC1 84.4%) than that explained by the fingerprints with the 1000 most informative bits. However, if the PCA analysis is performed after the removal of (i) the properties whose values are identical in more than 90% of the molecules and (ii) the molecular descriptors with correlation above 0.9, the explained variance falls (PC1 44.7%) and it becomes comparable to that obtained with the fingerprints. Moreover, Dendritic (PC1 39.6%) and Linear (PC1 47.2%) fingerprints explained more variance than that explained by the Radial (PC1 10.9%) and MOLPRINT2D (PC1 7.49%) fingerprints. In general, fingerprints bits are more orthogonal compared to the whole ordinary descriptors and perform better at explaining structure activity data.<sup>18</sup> [Table 1](#)

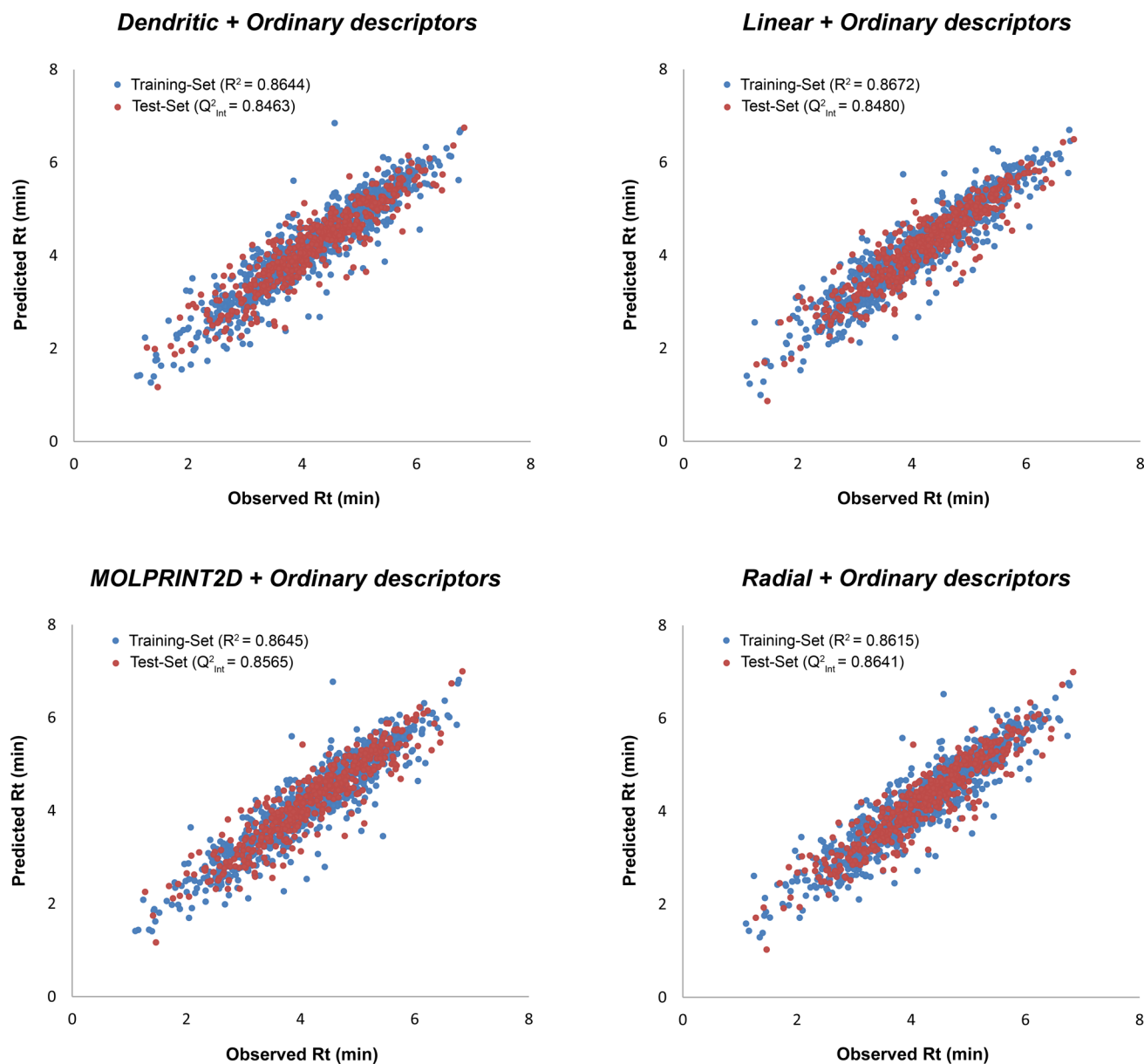
**Table 1. Predictive Power of the Best KPLS models**

model	$R^2$	$Q^2$ test set internal	$Q^2$ test set external
only traditional descriptors	0.8667	0.8488	0.8166
dendritic + ordinary descriptors	0.8644	0.8463	0.8271
linear + ordinary descriptors	0.8672	0.8480	0.8136
MOLPRINT2D + ordinary descriptors	0.8645	0.8565	0.8377
radial + ordinary descriptors	0.8615	0.8641	0.8303

summarizes the predictive power of the best KPLS models; further details are available in the [Supporting Information](#). A total of 10 classical multiple linear regression (MLR) models were built by changing the number of  $X$  variables from 1 to 10 (see [supplementary tables](#), MLR datasheet). With a number of  $X$  variables between 4 and 7, we obtained the best performance in terms of  $R^2$ ,  $Q^2$ , and predictive power on an external test set of newly synthesized compounds (see [Figures S-3–S-5](#)).

The retention of compounds in reversed-phase chromatography depends on the interactions between the analyte and the stationary phase.<sup>24</sup> These interactions include directional force, induction force, dispersion force, and hydrogen bond and can be related to the topological structures as well as the geometric and electronic features of the analyte. In our MLR models descriptors, the highest positive weights were the AlogP, the average valence connectivity index chi-1 (AvX1),<sup>25</sup> the Kier benzene-likeness index (KBLI),<sup>26</sup> whereas the highest negative weights were the HDB (number of hydrogen-bond donors) index (see the [Supporting Information](#), MLR models). Quite unsurprisingly, given the reversed-phase separation mode, the AlogP in each model was shown to represent an important contribution to the final equation. This is consistent with the fact that AlogP depends on the lipophilic character of the compound and, therefore, more lipophilic compounds are retained longer by a lipophilic stationary phase, as previously reported by other authors.<sup>4,17</sup> The average valence connectivity index chi-1 (AvX1) depends on the branched nature of the compound as well as on the presence and positions of heteroatoms. Therefore, compounds with higher connectivity interact more with the apolar stationary phase and their RT are higher. KBLI, describing the extent of molecular aromaticity, also correlates with RT, as a higher aromaticity helps the retention on the stationary phase. On the other hand, the



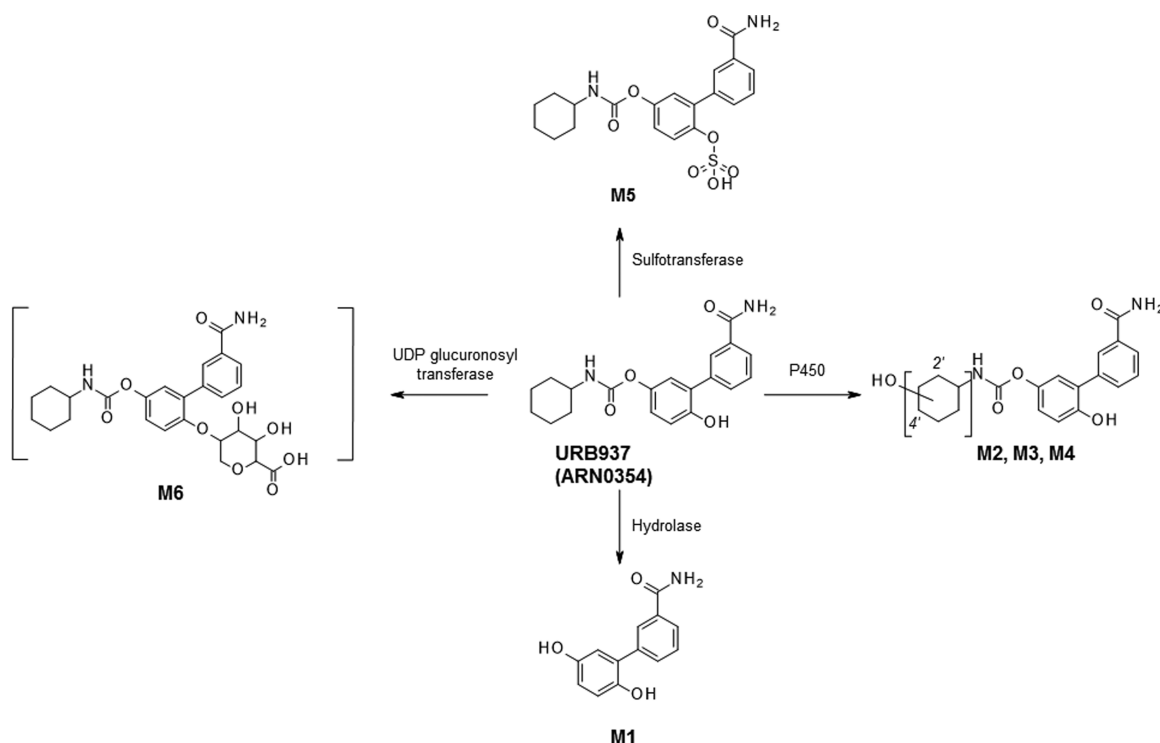


**Figure 1.** Calculated vs experimental RT values plots of the four combined (fingerprints + ordinary) KPLS regressions, for both Training (blue dots) and Test sets (red dots).  $R^2$  values are reported in the graph.

hydrogen bond forming capability, accounted for by the HBD descriptor, favors the interactions with the mobile phase, thus decreasing the RT of the compound. Finally, all the above MLR generated models were tested to predict the RTs of the metabolites of URB937 and showed a good predictive power ( $Q^2$  of 0.6923 for the model 7). Unfortunately, all the models underestimated the RT of the glucuronide derivative (see the [Supporting Information](#) (MLR model)).

**Performance Evaluation on an External Test Set.** A classical PLS analysis was conducted on the basis of the traditional descriptors with a maximum PLS factor of 5. Starting from a PLS factor of 3, the correlation factor  $R^2$  and the internal and external predictive power on a set of newly synthesized compounds were satisfactory (see PLS Table, [Supporting Information](#)). These linear methods, despite their good  $R^2$  and  $Q^2$  values, were not able to correctly predict the RT of URB937 metabolites. Therefore, nonlinear methods were preferred, even if interpretability issues may arise, as reported by Guha et al.<sup>27</sup>

Nine KPLS models were built, four models by using four fingerprint types (dendritic, linear, radial, and MOLPRINT2D), one on the basis of the ordinary descriptors and the other four by using a combination of the above fingerprints and the ordinary molecular descriptors. A maximum of five latent factors were used in each model, but this factor was automatically reduced when the standard deviation (SD) dropped below 0.4 to avoid overprediction. Several Kernel nonlinearity factors were tested (see [supplementary tables](#), PLS datasheet), with values close to zero corresponding to linear and large values to nonlinear. Higher nonlinearity usually leads to a tighter fitting, but it also gives poorer predictions on external compounds. Ten bootstrapping cycles were finally conducted to calculate the uncertainty for each attempt. A good model was obtained for each fingerprint set (see KPLS Table, [Supporting Information](#)). In particular, the best models were obtained by using dendritic and linear fingerprints with respect to the best models obtained with MOLPRINT2D and radial



**Figure 2.** Chemical structures of URB937 and observed Phase I and Phase II metabolites.

fingerprints. Dendritic and linear fingerprints performed similarly in terms of  $R^2$  but worse in terms of predictive power (internal  $Q^2$ ) with respect to the best model built with the ordinary descriptors. Interestingly, the combination of fingerprints and ordinary descriptors enhanced the performance of dendritic, MOLPRINT2D, and radial-fingerprint-derived models to values of  $R^2$  and  $Q^2$ , close to that of the model built with only the ordinary descriptors. Moreover, the model built with a combination of linear fingerprints and the ordinary descriptors slightly overperformed the  $R^2$  value of the best model built with the ordinary descriptors, while maintaining the same  $Q^2$  value. The results are reported in Figure 1. Our KPLS method also estimates the expected error in the predicted RT, calculated from 10 iterations of the algorithm. Following intrinsic control experiments performed on the learning set itself, we tested these four computational models on a new subset of 394 molecules (external test set), randomly selected from among the new chemical entities (NCEs) synthesized in our department in 2014. These NCEs were not present in the initial learning set and belonged to new chemical classes. RT values for these structures were predicted with the four models prepared using the learning set and plotted against their corresponding experimental values. For all the models, the obtained  $Q^2$  for the external data set ( $Q^2_{\text{ext}}$ ) exceeded 0.81 (see Table 1; supplementary tables, KPLS predictions data-sheet; and Figure S-3).

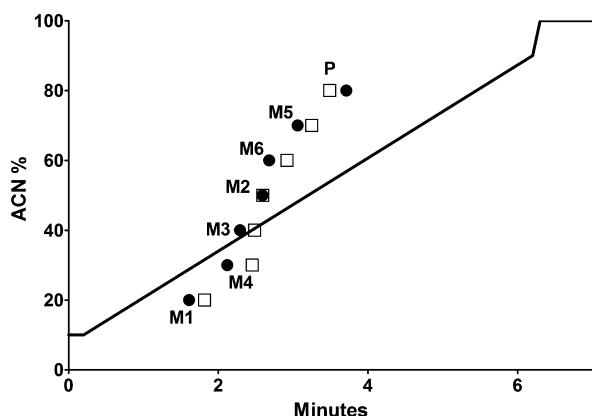
On the basis of these data, the Linear model performed best, with its predicted RTs fitting the experimental values with an  $R^2$  close to 0.87. As a comparison, it is worth reporting that a similar QSRR model<sup>12</sup> built on 120 metabolites for HILIC RT prediction shows a  $R^2$  value of 0.82. A detailed statistical analysis based on the obtained results for each model is reported in the Supporting Information. This analysis shows (a) the homogeneity of the variance and (b) that SDs are comparable with those obtained with experimental methods (as

order of magnitude). Canvas KPLS with Canvas fingerprints allow one to establish predictive and easy-to-interpret models with respect to other nonlinear (black box) methods.<sup>18</sup> Mixed KPLS models based on both fingerprints and classical descriptors, however, cannot be easily interpreted. On the other hand, although interpretability is crucial when structural modifications have to be hypothesized for improving compound's activity or reducing liabilities, in the case of our RT prediction model we preferred the accuracy of the prediction over the mechanistic interpretability of the LC behavior. Using this "black box" approach the obtained models returned satisfactory results. Finally, to evaluate this method's performance on a real-life analytical scenario, we challenged this algorithm with two LC–MS-based metabolite identification experiments, where a predicted RT value for isomeric species would be a valuable help in assigning chromatographic peaks to their correct structures.

**Application of the Model to Real-Life MetID Experiments.** As a first case study, we applied the model to the prediction of the RT of the major metabolites of URB937, a drug candidate for treating pain<sup>28,29</sup> (Figure 2). Phase I and Phase II metabolism of this compound produces six major metabolites that were carefully investigated by UPLC–MS/MS (see the Supporting Information for experimental details).

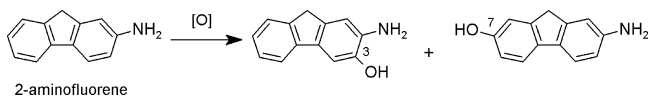
The three monohydroxylated isomeric species (M2, M3, and M4) were particularly hard to assign in our met ID experiments because their MS/MS fragmentation pattern did not allow a totally confident assignment of the 2', 3', and 4'-OH cyclohexyl isomers. As part of our preclinical studies on URB937, all the above-described metabolites were synthesized to achieve a conclusive structural elucidation by comparison with a standard reference, bringing additional and remarkable synthetic efforts to the project. We then considered this complex group of compounds as the ideal candidate for testing the robustness of our predictive model, given the added

difficulty of having six polar metabolites plus the parent eluting very close to each other in a short run, with a steep and nondedicated 10–90% ACN gradient in 6 min (a 15% B/min slope). All these molecules were thus analyzed under the standardized QC conditions. As expected, the nonoptimized ACN gradient makes five compounds elute in less than 1 min, with M2, M3, and M4 isomers eluting in only 25 s. Despite this, our KPLS algorithm performed reasonably well on this difficult challenge. Figure 3 reports the generic gradient profile used with the expected and experimental RT values for these seven compounds.



**Figure 3.** Predicted (□) vs experimental (●) RT values for URB937 metabolites. M1, hydroquinone; M4, 4'-OH; M3, 3'-OH; M2, 2'-OH cyclohexyl derivative; M6, glucuronide; and M5, sulfate adducts on the phenol group. P, parent compound.

As shown in Figure 3, our algorithm estimated the experimental elution time of all seven compounds with a relative error ranging from 0.4 to 15.7% of the observed value. The rapid gradient profile used for QC analysis makes the three mono-oxidized isomeric species elute close to one another, but our method correctly predicted their elution order as 4', 3', and 2'-hydroxy derivative with reasonable confidence. This data would have dramatically supported the metabolite identification task performed on URB937, without requiring additional synthetic chemistry efforts. As a second case of real-life metabolite ID experiment, we tested our method's ability to guess the elution order of the two major 2-aminofluorene (2AF) Phase I metabolites (Figure 4) in the same LC–MS conditions described above (15% B/min gradient).



**Figure 4.** Phase I oxidation of 2-aminofluorene into 3- and 7-hydroxy isomeric metabolites.

As already reported in the literature,<sup>30,31</sup> 2AF is metabolized by human liver enzymes into two major oxidized species, 3- and 7-hydroxy-2-aminofluorene, that being isomeric, are virtually impossible to distinguish by MS and MS/MS. Our KPLS model correctly guessed the elution order of the two species, as reported in Table 2. As for URB937, even with a nonoptimized gradient profile, a high confidence in metabolite ID is enabled by our KPLS model, because the predicted separation of the

**Table 2.** Experimental and Predicted RT Values (In Minutes) For 2-Aminofluorene (2AF) Metabolites

ID	experimental	predicted	difference	error %
7-OH, 2AF	2.53	2.70	0.17	6.7
3-OH, 2AF	3.42	2.91	0.51	14.9
2AF	4.16	3.67	0.49	11.7

two isomeric species (3-OH and 7-OH) is estimated in roughly 12 s.

**Extension to Other Gradient Profiles.** To make our KPLS model even more useful for metabolite ID, we then asked whether its predictive power could be extended to other chromatographic conditions. As reported by Eugster and colleagues in 2014,<sup>32</sup> if a different gradient profile is introduced, with no changes occurring in the stationary and mobile phases or in the column temperature (thermodynamic parameters), then the new RT value of a given compound can be calculated by changing the gradient slope and flow rate (kinetic parameters), using eq 1 to shift the RT values from an old to a new condition:

$$RT_{\text{new}} = (\%B_{i,\text{old}} - \%B_{i,\text{new}} + GS_1(RT_{\text{old}} - d_{\text{old}} - h_{\text{old}} - (dw_{\text{old}}/F_{\text{old}})))/GS_{\text{new}} + d_{\text{new}} + h_{\text{new}} + (dw_{\text{new}}/F_{\text{new}}) \quad (1)$$

RT values for a new method can be calculated by changing the kinetic parameters of an old separation. %B<sub>i</sub> = initial percentage of eluent B, GS = gradient slope, *d* = column dead time, *h* = initial isocratic hold, and *F* = flow rate.

We then recalculated the expected retention times for the whole data set (>1300 molecules) for a 5% per minute gradient using eq 1, and we reran our KPLS model to predict the retention times of URB937 metabolites. At the same time, we reran the compounds with this new gradient to collect their experimental RT values. The results are reported in Table 3.

As the table demonstrates, by simply cutting the slope steepness to a third (5% B/min), both the measured and the calculated separation between the compounds roughly triple, making the assignment of the three isomeric peaks even less ambiguous. In the most difficult case, the predicted M3/M4 peak separation is 0.03 min (less than 2 s) at 15%/min but it becomes 0.11 (a little less than 7 s) at 5%/min, thus allowing a confident peak assignment.

## CONCLUSIONS

We have here proposed a new QSRR model based on a kernel-based partial least-squares method for predicting UPLC retention times in reversed phase mode. The model was built using a combination of classical (physicochemical and topological) and not classical (fingerprints) molecular descriptors on a database of accurately measured retention time values from a 1383 compound library encompassing different chemical classes and structures. Following a random splitting of the data set into training set and test set, we assessed the ability of the model to predict the retention time of all the compounds. The best predicted/experimental *R*<sup>2</sup> value was higher than 0.86, while the best *Q*<sup>2</sup> value we observed was close to 0.84. A comparison of our model with traditional MLR and PLS regression models shows that KPLS better performs in term of correlation (*R*<sup>2</sup>), prediction (*Q*<sup>2</sup>) and support to MetID peak assignment. The KPLS model succeeded in two real-life

**Table 3. Changes in Experimental and Predicted RT Values (in Minutes) for URB937 Metabolites Moving the Gradient Slope from 15%/min (Actual Method) To 5% B/min<sup>a</sup>**

elution order	ID	15% B/min					5% B/min				
		measured	predicted	difference	error %	predicted separation	measured	predicted	difference	error %	predicted separation
M1	hydroquinone	1.61	1.82	0.21	13.00		2.20	2.95	0.75	34.10	
M4	4'-OH	2.12	2.45	0.33	15.60	0.63	3.93	5.03	1.10	28.00	2.08
M3	3'-OH	2.29	2.48	0.19	8.30	0.03	4.51	5.14	0.63	14.00	0.11
M2	2'-OH	2.59	2.60	0.01	0.40	0.12	5.52	5.50	0.02	0.40	0.36
M6	glucuronide	2.68	2.92	0.24	9.00	0.32	5.83	6.47	0.64	11.00	0.97
M7	sulfate	3.06	3.25	0.19	6.20	0.33	7.11	8.12	1.01	14.20	1.65
P	parent	3.71	3.49	0.22	5.90	0.24	9.32	8.62	0.70	7.50	0.50

<sup>a</sup>The predicted peak separation between a metabolite and the following is also reported as a distance from the preceding peak.

MetID tasks, by correctly predicting elution order of Phase-1 metabolites, including isomeric monohydroxylated compounds. We have also shown here that the model's predictive power can be extended to different gradient profiles, by simple mathematical extrapolation using a known equation, thus offering very broad applicability. Moreover, the current study includes a deep investigation of different types of chemical descriptors used to build the structure-retention relationships. The new QSRR model based on the kernel partial least-squares method is built on the experimental UPLC RT values of a large library of chemical compounds. Many chemical classes are represented in the vast learning set and therefore the model benefits from a high degree of chemical diversity. This molecular heterogeneity helps in removing any class-related bias, making the method suitable for a broad range of small organic molecules spanning over a large part of the chemical space of pharmacological interest.<sup>33</sup> For our purpose, following an extensive analysis of the current literature in the field and the tools available to support LC method development, we established our KPLS method testing different models. We then carefully conducted a statistical analysis of all of them, by splitting the data set in training and test set, and by carefully evaluating the model performances. All the models we tested showed a predicted vs experimental RT correlation ( $R^2$ ) higher than 0.86. We then undertook a mechanistic analysis of our predictive model, by carefully evaluating the molecular descriptors showing the highest and lower weights on the prediction. We then showed how this computational tool can greatly assist in correctly assigning isomeric structures, by reporting the results of two real-life metabolite identification experiments on Phase-1 metabolism of small molecules. Additional value of our method lies in the practical possibility of extending the model's predictive power from the standardized QC chromatographic system to any other gradient profile, greatly increasing the method's flexibility and applicability, by using already published RT extrapolation algorithm. Leaving aside the proposed method's intrinsic scientific and academic value, we note that the described protocol can be reproduced relatively easily by any pharmaceutical company with a library of compounds tested for purity by standardized UPLC and a KPLS algorithm. The generated algorithm will greatly assist in those metabolite identification experiments needed for DMPK profiling of the company's compounds, without the need of additional synthetic chemistry efforts, thus allowing consistent resource savings.

## ■ ASSOCIATED CONTENT

### § Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b02075.

Additional experimental procedures in detail and additional experimental data and analysis (PDF)

Detailed data analysis for the four KPLS models used (dendritic, linear, molprint, and radial) and for the better in term of prediction MLR model (XLSX)

Detailed data analysis for the dendritic KPLS model and for the better in terms of prediction MLR model (XLS)

Detailed data analysis for the linear KPLS model and for the better in terms of prediction MLR model (XLS)

Detailed data analysis for the molprint KPLS model and for the better in terms of prediction MLR model (XLS)

Detailed data analysis for the radial KPLS model and for the better in terms of prediction MLR model (XLS)

MLR, PLS, and KPLS models used and the KPLS predictive power (XLS)

## ■ AUTHOR INFORMATION

### Corresponding Author

\*E-mail: [andrea.armirotti@iit.it](mailto:andrea.armirotti@iit.it). Phone: +39 010 71781938. Fax: +39 010 71781228.

### Author Contributions

§F.F. and S.M.B. had equal contribution to this work.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

The authors wish to thank Dr. Silvia Venzano and Dr. Luca Goldoni for compound handling and support in QC analysis. We thank Prof. Daniele Piomelli for the scientific direction of the Drug Discovery and Development (D3) Department. We thank Grace Fox for copyediting and proofreading the manuscript.

## ■ REFERENCES

- (1) Novakova, L.; Solichova, D.; Solich, P. *J. Sep. Sci.* **2006**, *29*, 2433–2443.
- (2) Leandro, C. C.; Hancock, P.; Fussell, R. J.; Keely, B. J. *J. Chromatogr. A* **2006**, *1103*, 94–101.
- (3) Gika, H. G.; Macpherson, E.; Theodoridis, G. A.; Wilson, I. D. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* **2008**, *871*, 299–305.
- (4) Gorynski, K.; Bojko, B.; Nowaczyk, A.; Bucinski, A.; Pawliszyn, J.; Kaliszan, R. *Anal. Chim. Acta* **2013**, *797*, 13–19.



- (5) Andres, A.; Roses, M.; Bosch, E. *J. Chromatogr. A* **2014**, *1370*, 129–134.
- (6) Tyrkko, E.; Pelander, A.; Ojanpera, I. *Anal. Chim. Acta* **2012**, *720*, 142–148.
- (7) Hewitt, E. F.; Lukulay, P.; Galushko, S. *J. Chromatogr. A* **2006**, *1107*, 79–87.
- (8) Racz, N.; Kormany, R.; Fekete, J.; Molnar, I. *J. Pharm. Biomed. Anal.* **2015**, *108*, 1–10.
- (9) Baczek, T.; Kaliszan, R. *Proteomics* **2009**, *9*, 835–847.
- (10) Klammer, A. A.; Yi, X. H.; MacCoss, M. J.; Noble, W. S. *Lect. Notes Comput. Sc.* **2007**, *4453*, 459–472.
- (11) Krokhin, O. V.; Ying, S.; Cortens, J. P.; Ghosh, D.; Spicer, V.; Ens, W.; Standing, K. G.; Beavis, R. C.; Wilkins, J. A. *Anal. Chem.* **2006**, *78*, 6265–6269.
- (12) Creek, D. J.; Jankevics, A.; Breitling, R.; Watson, D. G.; Barrett, M. P.; Burgess, K. E. *Anal. Chem.* **2011**, *83*, 8703–8710.
- (13) Bade, R.; Bijlsma, L.; Sancho, J. V.; Hernandez, F. *Talanta* **2015**, *139*, 143–149.
- (14) Barron, L. P.; McEneff, G. L. *Talanta* **2016**, *147*, 261–270.
- (15) Miller, T. H.; Musenga, A.; Cowan, D. A.; Barron, L. P. *Anal. Chem.* **2013**, *85*, 10330–10337.
- (16) D'Archivio, A. A.; Maggi, M. A.; Ruggieri, F. *J. Pharm. Biomed. Anal.* **2014**, *96*, 224–230.
- (17) Aalizadeh, R.; Thomaidis, N. S.; Bletsou, A. A.; Gago-Ferrero, P. *J. Chem. Inf. Model.* **2016**, *56*, 1384–1398.
- (18) An, Y.; Sherman, W.; Dixon, S. L. *J. Chem. Inf. Model.* **2013**, *53*, 2312–2321.
- (19) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. *J. Chem. Theory Comput.* **2016**, *12*, 281–296.
- (20) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. *J. Chem. Inf. Model.* **2010**, *50*, 771–784.
- (21) Moreno-Sanz, G.; Duranti, A.; Melzig, L.; Fiorelli, C.; Ruda, G. F.; Colombano, G.; Mestichelli, P.; Sanchini, S.; Tontini, A.; Mor, M.; Bandiera, T.; Scarpelli, R.; Tarzia, G.; Piomelli, D. *J. Med. Chem.* **2013**, *56*, 5917–5930.
- (22) Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A. *J. Chem. Inf. Model.* **2012**, *52*, 2570–2578.
- (23) Falchi, F.; Manetti, F.; Carraro, F.; Naldini, A.; Maga, G.; Crespan, E.; Schenone, S.; Bruno, O.; Brullo, C.; Botta, M. *ChemMedChem* **2009**, *4*, 976–987.
- (24) Braumann, T. *Journal of chromatography* **1986**, *373*, 191–225.
- (25) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure–Activity Analysis*; Research Studies Press: Letchworth, Hertfordshire, England, 1986.
- (26) Asadollahi-Baboli, M. *SAR QSAR Environ. Res.* **2012**, *23*, 505–520.
- (27) Guha, R. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 857–871.
- (28) Sasso, O.; Bertorelli, R.; Bandiera, T.; Scarpelli, R.; Colombano, G.; Armirotti, A.; Moreno-Sanz, G.; Reggiani, A.; Piomelli, D. *Pharmacol. Res.* **2012**, *65*, 553–563.
- (29) Clapper, J. R.; Moreno-Sanz, G.; Russo, R.; Guijarro, A.; Vacondio, F.; Duranti, A.; Tontini, A.; Sanchini, S.; Sciolino, N. R.; Spradley, J. M.; Hohmann, A. G.; Calignano, A.; Mor, M.; Tarzia, G.; Piomelli, D. *Nat. Neurosci.* **2010**, *13*, 1265–1270.
- (30) Hammons, G. J.; Guengerich, F. P.; Weis, C. C.; Beland, F. A.; Kadlubar, F. F. *Cancer research* **1985**, *45*, 3578–3585.
- (31) Stanley, L. A.; Skare, J. A.; Doyle, E.; Powrie, R.; D'Angelo, D.; Elcombe, C. R. *Toxicology* **2005**, *210*, 147–157.
- (32) Eugster, P. J.; Boccard, J.; Debrus, B.; Breant, L.; Wolfender, J. L.; Martel, S.; Carrupt, P. A. *Phytochemistry* **2014**, *108*, 196–207.
- (33) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Delivery Rev.* **2001**, *46*, 3–26.