

"This accepted author manuscript is copyrighted and published by Elsevier. It is posted here by agreement between Elsevier and MTA. The definitive version of the text was subsequently published in [JOURNAL OF CHROMATOGRAPHY A, volume 1380: 130-138 (2015) [doi:10.1016/j.chroma.2014.12.073](https://doi.org/10.1016/j.chroma.2014.12.073)]. Available under license CC-BY-NC-ND."

Chromatographic and computational assessment of lipophilicity using sum of ranking differences and generalized pair-correlation

Filip Andrić¹, Károly Héberger^{2,*}

¹Faculty of Chemistry, University of Belgrade, Studentski trg 12-16,

²Research Centre for Natural Sciences, Hungarian Academy of Sciences, H-1117 Budapest XI., Magyar Tudósok krt 2.

*Corresponding author: Károly Héberger

Phone: +36 1 38 26 509

E-mail: heberger.karoly@ttk.mta.hu

Abstract

Lipophilicity ($\log P$) represents one of the most studied and most frequently used fundamental physicochemical properties. At present there are several possibilities for its quantitative expression and many of them stems from chromatographic experiments. Numerous attempts have been made to compare different computational methods, chromatographic methods vs. computational approaches, as well as chromatographic methods and direct shake-flask procedure without definite results or these findings are not accepted generally.

In the present work numerous chromatographically derived lipophilicity measures in combination with diverse computational methods were ranked and clustered using the novel variable discrimination and ranking approaches based on the sum of ranking differences and the generalized pair correlation method. Available literature $\log P$ data measured on HILIC, and classical reversed-phase combining different classes of compounds have been compared with most frequently used multivariate data analysis techniques (principal component and hierarchical cluster analysis) as well as with the conclusions in the original sources. Chromatographic lipophilicity measures obtained under typical reversed-phase conditions outperform the majority of computationally estimated $\log Ps$. Oppositely, in the case of HILIC none of the many proposed chromatographic indices overcomes any of the computationally assessed $\log Ps$. Only two of them ($\log k_{\min}$ and k_{\min}) may be selected as recommended chromatographic lipophilicity measures. Both ranking approaches, sum of ranking differences and generalized pair correlation method, although based on different backgrounds, provides highly similar variable ordering and grouping leading to the same conclusions.

Keywords: Lipophilicity, Multivariate data analysis, Sum of ranking differences, Generalized pair correlation method, high-performance liquid chromatography

1. Introduction

Since the first works of Meyer and Overton [1,2] the term lipophilicity was tailored to what is now used as a fundamental physicochemical property in many quantitative structure-activity relationships (QSAR), quantitative structure-property relationships (QSPR) studies, pharmaceutical and environmental sciences, as well as toxicological assessments [3].

Modern definition of lipophilicity according to the International Union for Pure and Applied Chemistry (IUPAC) is as follows: lipophilicity represents the affinity of a molecule or a moiety for a lipophilic environment. The same definition also describes shortly the methods for its measurements – “it is commonly measured by its distribution behavior in a biphasic system, either liquid-liquid (*e.g.*, partition coefficient in 1-octanol/water) or solid-liquid (retention on reversed-phase high-performance liquid chromatography (RP-HPLC) or thin-layer chromatography (TLC) system)” [4].

At present there are several means for quantitative expression of lipophilicity, and many of them have been derived from chromatographic experiments [5-8]. However, the most commonly used is the octanol-water partition coefficient. It is defined as the ratio of the concentrations of a neutral compound in octanol and aqueous phases that are under equilibrium conditions. Different notations can be found in literature such as: $\log P_{O/W}$, $\log K_{OW}$ or just simply $\log P$. For the sake of simplicity in this paper the last term will be applied.

Several experimental techniques for measuring $\log P$ have been developed so far and some of them are implemented as the standard tests through the guidelines of the Organization for Economic Cooperation and Development (OECD) such as: Test No. 107, Shake flask method [9], Test No. 117, HPLC method [10], and Test No. 123, Slow stirring method [11]. However, depending on the method used, experimental determination is linked to numerous difficulties such as formation of stable emulsion between *n*-octanol and water.

Classical shake-flask approach is time and reagent consuming and it is also unsuitable for impure compounds or compounds of extremely low or high $\log P$ values ($-3 < \log P < 4$). Therefore, the development of simpler, yet more accurate experimental methods is a valuable aim. Chromatographic methods have several advantages. They are easy to employ, they give coherent results in the similar $\log P$ range as the shake-flask method, and contaminated or degraded compounds can be analyzed as well. Also, interactions that are responsible for retention of a solute can be tuned in such a way to get as much as possible of its lipophilic character, simply by selecting appropriate chromatographic conditions. Reversed-phase modalities that utilize highly non-polar stationary phase such as various hydrocarbon modified silica gels (octadecyl, octyl, ethyl, and phenyl commonly denoted as C18, C8, C2 and Ph respectively) [12], or amphiphilic sorbents such as cyano-propyl, amino-propyl or diol modified silica [13] in combination with polar mobile phase (water - organic solvents mixtures) are usually employed. Hydrophilic interaction liquid chromatography (HILIC) [14] or salting-out chromatography [15] may be a good choice for analysis of highly polar solutes.

Both, HPLC and TLC provide almost endless series of chromatographic descriptors that can be used for quantitative expression of lipophilicity [16,17]. They are derived either directly from retention data or extrapolated from linear relationships between retention and mobile phase composition. Short summary of chromatographic lipophilicity measures along with brief description and chromatographic modality is given in the **Table 1**. Many of these properties have been so far extensively used in QSPR, QSAR studies [18-21].

Table 1

Although experimentally determined values are preferred, computational approaches have significant advantages because they do not require expensive instrumentation, reagents and laborious experimental work. Also, their extensive use emerges from demands of industry on

fast, simple, high throughput, and yet reliable ways to provide information needed for fast screening of target compounds. So far many calculation techniques have been developed and basically all can be classified in two major groups: substructure-based and property-based approaches. Substructure-based approaches simply split the structure into fragments (fragment-based) or even down to the level of atoms (atom-based). Then all substructure contributions are added up using contribution terms and correction factors to obtain the final $\log P$ values. Property-based approaches, on the other hand, utilize descriptions of the molecule as a whole. They are based either on empirical approaches such as linear solvation energy relationships (LSER), or 3D-structure representation (COSMOFrag), models that utilize topological, electrotopological, or simple 1D descriptors (AlogPs, MLOGP). The computationally estimated $\log P$ scales that are used in this paper are summarized in Table 2. However, still various calculation methods provide 2-3 order of magnitude difference in $\log P$ values for the same molecule, which might question the reliability of these methods on a large scale.

Table 2

Lipophilicity strongly affects compound solubility, passive transport through biological membranes including gastrointestinal absorption, blood-brain barrier, drug-receptor binding influencing bioavailability, biodistribution, toxicity, including ecotoxicity, *etc.* Hence, the choice of appropriate lipophilicity measures is crucial for modeling biological response, as well as environmental processes.

So far there have been many attempts to compare chromatographic methods *versus* computational approaches [22,23-27], and to compare different computational methods [24,25], as well as chromatographic methods and direct shake-flask procedure. Such comparisons mainly relied either on establishing good correlations, *i.e.*, mathematical models between the studied properties, using parametric statistical parameters as a quality measure,

namely: Pearson's correlation coefficient, various error estimations and measures of model predictive power or they relied on the multivariate exploratory analysis such as principal component analysis (PCA) and hierarchical cluster analysis (HCA) to search for possible similarities among studied lipophilicity measures. However, even parametric quantities such as aforementioned correlation coefficients and error estimates are inferior when it comes to decisions based on slight differences.

Therefore, the aim of the present work was to provide better understanding, and give a critical review of relations among different chromatographic modalities and various chromatographically derived lipophilicity measures in combination with $\log P$ computational methods. Our aim was also to rank and group the lipophilicity measures, to select the best and worse one and to give recommendations about their usage (or whether their usage should be avoided). Therefore, we employed the novel comparison/ranking approaches based on the sum of ranking differences (SRD) and the generalized pair correlation method (GPCM).

2. Materials and methods (Calculations)

2.1. Collection of lipophilicity data

Lipophilicity data have been collected from two sources that will further be presented in a form of two case studies (see sections 3.1. and 3.2.). Literature has been selected in such a way so that different chromatographic modalities such as typical reversed-phase conditions [22] or hydrophilic interaction chromatography [23] are covered. Also, significant diversity of chromatographically derived lipophilicity indices and computational approaches to calculation of $\log P$ values has been taken into account as well.

2.2. Data pretreatment and exploratory statistical analysis

Because the data were comprised of variables of different nature they are expressed on the same scale using: (i) standardization (mean centered and scaled to unit standard deviation), (ii) range scaling between lowest and highest computationally estimated $\log P$ value and (iii) rank transformation. Autoscaled data were used for the exploratory data analysis employing unsupervised classification techniques, HCA and PCA, while the interval scaled and ranked data have been used together with the autoscaled ones for the variable comparison by means of SRD and GPCM. Euclidian distance as the measure of dissimilarity was applied in case of HCA. Ward's method was used to define the distance among groups (linkage rule). PCA has been performed using PCA and multivariate/Batch SPC module as a part of Statistica v. 10 (Statsoft Inc. Tulsa, Oklahoma, USA). The significant number of principal components has been determined based on the scree plot. Other data treatments were done using MS Microsoft Excel 2010.

2.3. Sum of ranking differences and comparison with random numbers

Sum of ranking differences was introduced in the field of analytical chemistry as a method that fairly compares methods or models [29]. It has been already applied in different fields, *e.g.*, column selection in chromatography [29], comparison of predictive performance of QSAR models [29,30]; selection of the best polarity measure for small organic molecules [31]; for testing panel consistency in food chemistry [32], for comparison of various of comet assay parameters for genotoxicity testing [33], *etc.* Detailed theoretical basis of SRD method is given elsewhere [29,34,35]. However, some basic principles are worth to be described here. The method is entirely general and supervised in the sense that it requires some benchmark or a reference ranking. The objects and variables, in our case different compounds and lipophilicity measures, are arranged in a form of a matrix, *i.e.*, in rows and columns respectively. There are essentially two possibilities to choose benchmark values. The first one

is to make average value from all data in one row (so called row-average), and do that for all rows (objects). The other possibility is to choose the reference values, for example $\log P$ values measured by the shake-flask method. The first approach, called consensus, is in accordance with the maximum likelihood principle, which yields a choice of the estimator as the value for the parameter that makes the observed data most probable (the average). All methods (variables) have some random errors that cancel each-other using the average. It is a well-substantiated empirical finding in analytical chemistry that the systematic errors (biases) of different laboratories (measurement methods) follow normal distribution and hence they also cancel each other. However, the average is not the only option for data fusion.

In this work, all lipophilicity scales after being pretreated and expressed on the same scale were further ranked and compared with the ranked benchmark (average). Then, absolute differences of ranks between benchmark and each variable were calculated for every single molecule and then summed into SRD value(s). The closer the SRD value is to zero, the better is that particular variable (lipophilicity measure). Also, the mutual proximity of SRD values indicates the specific grouping of variables.

Validation of the SRD procedure was completed in two ways: The first approach, called comparison of ranks by random numbers (CRRN), either uses simulated random numbers or theoretical distribution of the random SRD values as described in ref. [36]. Calculated SRD values that significantly differ from random distribution fall away from each side of the theoretical or fitted Gaussian curve at the probability level $p = 0.05$.

The other way of validation is a seven fold cross-validation. Namely, approximately 1/7 of objects are omitted and the ranking is performed on the remaining data set. In that way seven SRD values are produced for each variable and the standard deviation is calculated, providing the insight into variability of every particular variable. Statistical difference among variables can be tested by applying Wilcoxon's matched pair test, as well as sign test on the seven SRD

values for each pair of variables. Also, an overview of uncertainties (distribution) for variables is done in a form of box and whisker plot according to the following criteria: 1) increasing median values, if the median values 2) are the same then the quartiles and interquartile ranges are taken into account. The first and third quartiles of two variables have the same “power”: if the two first quartiles (for two variables) are the same then, the smaller 3rd quartile should be the first. If the two 3rd quartiles of two variables are the same then, the smaller 1st quartile should be the first. If they are contradictory, then and only then the larger interquartile range counts. If they are equal, then 3) the maximum and minimum of two consecutive variables are checked. If the two minima are the same, then the smaller maximum should be the first. If the two maxima are the same then, the smaller maximum should be the first. If they are contradictory, then and only then the larger range counts between minimum and maximum of two variables. If they are all equal, no decision can be made. Box and whisker plots provide additional insight into grouping of variables and their statistical significance.

2.4. General pair correlation method (GPCM)

GPCM approach is based on completely different background than SRD. The procedure is already described in detail [36,37]. Basically, the method compares variables pair-wise in all possible combinations. Any of the two variables are compared to the benchmark variable and decided, which one is superior, inferior, or no decision can be made. Frequencies of wins, losses, *etc.* are counted. A few statistical tests may be used to determine statistical significance of decision but in the present work only Conditional exact Fisher’s test has been used.

Furthermore, all variables are ranked according to the probability weighted wins minus losses, *i.e.*, $p(\text{wins})-p(\text{losses})$ scores, which have been further reversely scaled in order to be

comparable with the SRD values. Arithmetic means of all rows were chosen as the benchmark.

3. Results and discussion

3.1. Case study 1 Comparison of calculated lipophilicity measures with the measured ones by HPLC - hydrophilic interaction chromatography

The data for this case study have been obtained from the Table 1 and the supplementary material Part 2 of ref. [23]. The authors measured retention of 30 solutes, pyridinium oximes, therapeutically tested in acetylcholinesterase reactivation, under bimodal chromatographic conditions, *i.e.*, reversed-phase and hydrophilic liquid chromatography, using HPLC technique. They provided 14 lipophilicity chromatographic indices for charged molecules, namely: k_{\min} , $\log k_{\min}$, ISOELUT, LOGISOELUT, ISOELUT1, LOGISOELUT1, ISOELUT2, LOGISOELUT2, $k_{\text{w}}^{\text{lin}}$, $\log k_{\text{w}}^{\text{lin}}$, $k_{\text{w}}^{\text{bin}}$, $\log k_{\text{w}}^{\text{bin}}$, HYL, and LOGHYL. These properties were compared with eleven computationally estimated lipophilicity scales: ALOGPs, AClogP, miLogP, KOWWIN, XLOGP2, XLOGP3, Hy, MLOGP, ALOGP, logD7, SlogD7.4. Some of the chromatographic lipophilicity indices as well as computationally estimated scales have been already listed in Table 1, however in addition to that, $k_{\text{w}}^{\text{lin}}$ and $k_{\text{w}}^{\text{bin}}$ represent retention factors extrapolated to the pure water content (zero content of the mobile phase organic modifier) using linear (lin) or binomial (bin) calculation approach. The authors introduced novel chromatographic lipophilicity measures such as: k_{\min} , $\log k_{\min}$, ISOELUT, LOGISOELUT, LOGISOELUT1, LOGISOELUT1, ISOELUT2, LOGISOELUT2, stating that some of them (k_{\min} , $\log k_{\min}$, ISOELUT, LOGISOELUT) are better correlated with calculated $\log P$ values than the rest of the lipophilicity indices.

In order to perform multivariate exploratory analysis and comparison by means of the SRD and GPCM, the data were arranged in a matrix form containing 30 rows (studied compounds

– objects) and 28 columns (lipophilicity measures – variables) and they were pretreated according to the described procedures. The entire dataset can be found in supporting information, Tables S1a and S1b (supplementary material).

3.1.1. Multivariate exploratory and classification analysis

In order to reveal the presence of any outliers, similarities and grouping patterns among variables, PCA and HCA were performed. PCA resulted in two principal components that account for the majority of the data variability, *i.e.*, *PC1* for 69.44% and *PC2* for 11.23, in total 80.67%. We obtained similar results as the authors have already reported [23], with one difference, we had to multiply the hydrophobicity descriptor (*Hy*) by -1, in order to be directly proportional to the $\log P$ values, and therefore positioned in the proximity of other lipophilicity parameters in the PCA loading plots, as appropriate.

Two major groups of variables and three possible outliers are present in *PC1/PC2* loading plot (Fig. 1). The first group consist of lipophilicity indices derived from the hydrophilic interactions dominant part of HILIC U-shape retention profile: ISOELUT1, ISOELUT2, LOGISOELUT1, LOGISOELUT2, HYL, and LOGHYL, while the second group consists of strongly overlapped computationally estimated $\log P$ values and several chromatographic lipophilicity measures obtained mostly from the reversed-phase end of U-shape retention curve, with exception of ISOELUT, LOGISOELUT, k_{\min} , and $\log k_{\min}$. The following: $PC1/k$, $PC1/\log k$, and $\log k_{\text{w}}^{\text{bin}}$ can be considered as outliers to the first group. Lipophilicity indices k_{\min} and $\log k_{\min}$ are located in the very heart of the second group.

Similar pattern is observed in the case of HCA (Fig. 2). Two clusters are formed above ten linkage distance units. Most chromatographic descriptors measured under HILIC modality are part of the first cluster, with exception of $\log k_{\text{w}}^{\text{bin}}$, while the second cluster combines computationally estimated ones as well as several reversed-phase chromatographically

determined parameters. Both, k_{\min} , and $\log k_{\min}$ are also tightly bound into single sub-cluster together with ISOELUT and LOGISOELUT parameters. Also, computational $\log P$ s exhibit higher degree of similarity (the smallest linkage distance) compared to chromatographic descriptors. Similarity measures based on correlation coefficient may reveal that among all studied variables only k_{\min} , and $\log k_{\min}$ are exceptionally well correlated (average $R^2 = 0.9332$ and 0.9423 , respectively) with majority of computationally estimated $\log P$ values.

Fig. 1

Fig. 2

3.1.2. Comparison of lipophilicity scales by SRD and GPCM

Although an inspection of PCA and HCA plots leads to several conclusions about similarities among studied lipophilicity scales, it is still impossible to choose, which one represents the best lipophilicity measure. Also, the information provided by correlation coefficients might lead to questionable conclusion that k_{\min} , and $\log k_{\min}$ could be the most suitable lipophilicity measures simply because they are best correlated with computationally estimated lipophilicity scales, especially since the rest of variables, except LOGISOELUT2 and $\log k_w^{\text{bin}}$ are statistically significantly correlated (Table S2, supplementary material) as well. Therefore, the use of non-parametric, robust methods that are able to compare, group, and rank variables such as SRD and GPCM is necessary in this case.

In order to apply SRD procedure it is mandatory that all variables should be put on the same scale. This was done by autoscaling, range scaling and ranking.

Fig. 3

According to the SRD-CRRN, the best lipophilicity measures, the closest to the zero value, are computationally estimated AClogP (in the case of autoscaled data) (Fig. 3), and XLOGP3 in the case of range scaled and ranked data (Table S5, supplementary material). These are followed by SlogD7.4, ALOGP, ALOGPs, $\log k_{\min}$, k_{\min} , SlogP, MLOGP,

KOWWIN and XLOGP2. Both, $\log k_{\min}$, and k_{\min} fall in this group. The order of lipophilicity scales is slightly altered depending of the data pretreatment; however, similarities in ranking are obvious. In addition, LOGHYL, HYL, and LOGISOELUT2 are not significantly different from random number distribution. Therefore they may be considered as unsuitable lipophilicity measures (not recommended variables). The rest of chromatographically derived variables fall between the mentioned categories, and follow more or less the same order.

Compared with approach based on correlations among computationally and chromatographically estimated lipophilicity measures, employed by Voicu *et al.* [26], SRD is more sensitive in separating non-significant variables (HYL, and LOGHYL in addition to LOGISOELUT2).

In addition to SRD-CRRN, SRD ranking based on sevenfold cross-validation and GPCM ranking are performed (Fig. 4 and 5). In the first case lipophilicity measures are arranged in increasing order of medians of SRD values. Both comparison methods share similar patterns with the corresponding SRD-CRRNs. In that sense, only slight differences can be noticed between them when applied on autoscaled and ranked data (Table S5, supplementary material). Interval scaled data exhibit some higher level of discrepancy. However, all milestone variables such as: k_{\min} , $\log k_{\min}$, ISOELUT, LOGISOELUT, $k_{\text{w}}^{\text{bin}}$, $\log k_{\text{w}}^{\text{bin}}$, HYL, LOGHYL, LOGISOELUT2, follow the same order (Table S5, supplementary material). This similarity is reassuring because the two methods (SRD and GPCM) have entirely different theoretical background and way of calculation.

Fig. 4 and Fig. 5

Finally, based on all three comparison methods none of the chromatographically derived lipophilicity indices outperform the computationally estimated lipophilicity measures. However, the best chromatographic lipophilicity descriptors are $\log k_{\min}$ and k_{\min} . They are in the first groups of ‘good’ lipophilicity descriptors (similar conclusion was provided by Voicu

et al.). According to the groupings of the sevenfold cross-validation ISOELUT2 separates the acceptable and the not recommended descriptors.

The following parameters can be considered as the best lipophilicity measures: SlogD7.4, XlogP3, AClogP, ALOGP, ALOGPs, k_{\min} , $\log k_{\min}$, SlogP and KOWWIN. The most unsuitable lipophilicity measures are: $PC1/\log k$, $\log k_{\text{w}}^{\text{bin}}$, ISOELUT1, HYL, LOGISOELUT2, and LOGHYL. Naturally, the ranking will be valid only for the given set of compounds using the given set of variables.

3.2. Case study 2 Comparison of calculated lipophilicity measures with the measured ones by reversed-phase liquid chromatography

In this case we have chosen the data from the Tables 1 and 2 of ref. [22]. The authors measured retention data of 23 flavonoids (neutral molecules) under the typical reversed-phase conditions using highly end-capped octadecyl silica (C18), polar embedded linker octadecyl silica (SB-18 Aqua), phenyl silica and pentafluorophenyl modified silica (PFP) as stationary phases. Chromatographic experiments were carried out by isocratic elution with acetonitrile-water mixtures at different volume fraction ratios. Several chromatographic lipophilicity measures were used: $\log k_{\text{w}}$, $m\log k$, S , ϕ_0 and $PC1/\log k$ and compared with 19 computational $\log P$ calculation methods. The authors also completed a PCA to study similarities and dissimilarities among the stationary phases. They also reported statistically significant correlations among calculated and chromatographically estimated lipophilicity scales.

3.2.1. Multivariate exploratory and classification analysis

Principal component analysis resulted in two components describing 86.59% of the overall data variability ($PC1$ - 74.85%, and $PC2$ - 11.74%) (Fig. 6). There is good separation among chromatographic and computationally estimated data along the $PC2$ axis (red line

through the origin: positive range - group A, negative range - group B). Also, $S(\text{C18}')$ and $\log k_w^{\text{bin}}(\text{C18})$ have low $PC1$ loading values. Since almost 75% of variability of lipophilicity data have been described by $PC1$ this should imply that these properties are not the most suitable lipophilicity measures.

Fig. 6

Fig. 7

HCA gives similar grouping (Fig. 7). Two clusters, A and B are formed at the linkage distance of around 13 and above. First one is mainly composed of chromatographic lipophilicity indices, with specifically separated $S(\text{PhF5})$, $S(\text{Ph})$ and $S(\text{C18}')$, while the other one includes predominately computationally estimated $\log P$ scales. However, neither PCA nor HCA do provide sufficient information regarding the most suitable lipophilicity measures.

3.2.2. Comparison of lipophilicity scales by SRD and GPCM

According to the SRD-CRRN, the typical reversed-phase mode provides lipophilicity indices that are more suitable in describing lipophilic characteristics of the studied compounds than HILIC, as expected. The best descriptors are obtained using C18 and C18' stationary phases and the best performances have ϕ_0 and $PC1/\log k$, which are followed by $\log k_w$ or $m\log k$ (Fig. 8). Lower ranking values were obtained in the case of descriptors measured on phenyl modified as well as pentafluorophenyl-modified silica. Vast majority of chromatographic indices are better ranked than computational methods (two separate variable categories can be distinguished, the first one with SRD values below 20 % comprising almost only chromatographic descriptors, while the second one with SRD values between 20 – 30 % consisting of mostly computational $\log P$ values). Slopes $S(\text{PhF5})$, $S(\text{C18})$, $S(\text{Ph})$ and $S(\text{C18}')$ can be considered as the worst lipophilicity measures, while $S(\text{C18}')$ together with $\log k_w^{\text{bin}}(\text{C18})$ do not differ significantly ($p = 0.05$) from the random number distribution.

These two parameters also do not show significant correlations ($p = 0.05$) with computationally estimated $\log P$ parameters, *i.e.*, $\log k_w^{\text{bin}}(\text{C18})$ show complete absence of correlation, while $S(\text{C18}')$ correlate poorly but significantly with $\log Da$, $ABlogP$, and COSMOFrag (Table S4, supplementary material).

Fig. 8

A sevenfold cross-validation procedure was employed to reveal the significance in the ordering. Similar pattern of ranked variables was obtained (Fig. 9). The lowest SRD values (10-20 %) were obtained for chromatographic lipophilicity descriptors, φ_0 , $PC1/\log k$ and $m\log k$, with particular ordering of stationary phases $\text{C18} < \text{C18}' < \text{Ph} < \text{PhF5}$. The variability of most chromatographic data is lower compared to the computationally calculated values (see lower and upper interquartile ranges in the box and whisker plot).

Fig. 9

GPCM ranking, although being completely different methodology, shows considerable similarity with the ordering and grouping of SRD-CRRN, with few exceptions (Fig. 10). First, numerous degeneration of variables occurs (if a methodology cannot distinguish variables (lipophilicity parameters) we call it degeneration or degeneracy). Second, all variables can be roughly divided just into two categories. Variables that fall into the first part of the graph (scaled $[p(\text{wins}) - p(\text{losses})] < 30$) are mostly chromatographic descriptors, while in the second part (scaled $[p(\text{wins}) - p(\text{losses})] > 30$) are composed from both. In addition, the slopes $S(\text{Ph})$, $S(\text{C18})$, $S(\text{PhF5})$, and $S(\text{C18}')$ possess high rank values. Despite of some differences among GPCM and SRD, all milestone variables follow similar order and grouping (φ_0 , $PC1/\log k$, $\log k_w$ and $m\log k$) or the type of stationary phase ($\text{C18} < \text{C18}' < \text{Ph} < \text{PhF5}$). MLOGP separates the best (and recommended) descriptors from the remaining ones (the same can be seen in Fig. 8-9). Naturally, the ranking will be valid only for the given set of compounds using the given set of variables.

No specific ordering of computationally assessed $\log P$ values is observed.

4. Conclusions

Both non parametric procedures, SRD-CRRN and GPCM, in both case studies give very similar results. In the case of hydrophilic interaction chromatography only few chromatographic parameters have been proven to have the most descriptive power as the computationally estimated $\log P$ methods, namely: $\log k_{\min}$ and k_{\min} , which are closely followed by ISOELUT and LOGISOELUT. In this particular case, classical chemometric methods (PCA loading plots as well as HCA analysis) support the SRD and GPCM ranking and grouping. In the case of reversed-phase HPLC majority of chromatographic descriptors outperforms most of the computationally assessed $\log P$ measures. In the first case the best lipophilicity measures are φ_0 , $PC1/\log k$, $\log k_w$ and $m\log k$, and the most suitable stationary phases follow the order $C18 > C18' > Ph > PhF5$.

In both case studies no specific pattern, ordering, or grouping of computationally estimated $\log P$ parameters according to the employed approach of computation, atom-based, fragment-based, mixed, or property based, is observed. Comparing SRD evaluation with GPCM, the latter has more degeneracy, *i.e.*, in some cases GPCM cannot distinguish the lipophilicity parameters whereas SRD and its cross-validated version can.

5. Acknowledgments

This work has been supported by the Ministry of Education, Science and Technological development, of the Republic of Serbia, grant No. 172017. KH is indebted to the financial support from OTKA, contract No. K112547.

6. References

- [1] H. Meyer, Zur Theorie der Alkoholnarkose I. Welche Eigenschaft der Anästhetika bedingt ihre narkotische Wirkung?, Arch. Pharmacol. Exp. Pathol. 42 (1899) 109-118.
- [2] E. Overton, Studien über die Narkose, zugleich ein Beitrag zur allgemeinen Pharmakologie, (1901) Jena, Switzerland: Gustav Fischer.
- [3] Zivoslav Lj. Tesic and Dusanka M. Milojkovic-Opsenica, TLC determination of Drug Lipophilicity chpt. in the book Thin Layer Chromatography in Drug Analysis, Edt. Lukasz Komsta, Monika Vaksmundzka-Hajnos, Joseph Sherma, (2013) CRC Press, Taylor & Francis Group.
- [4] IUPAC, Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"), Compiled by A. D. McNaught and A. Wilkinson, (1997) Blackwell Scientific Publications, Oxford. XML on-line corrected version: <http://goldbook.iupac.org> (2006) created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins
- [5] F. Murakami, Retention behavior of benzene derivatives on bonded reversed-phase columns, J. Chromatogr. 178 (1979) 393-399.
- [6] M. L. Bieganska, A. Doraczynska-Szopa, A. Petruczynik, The retention behavior of some sulfonamides on different TLC plates. 2. Comparison of the selectivity of the systems and quantitative determination of hydrophobicity parameters, J. Planar Chromatogr. – Mod. TLC 8 (1995) 122-128.
- [7] C. Sarbu, T. Sorina, Determination of lipophilicity of some non-steroidal anti-inflammatory agents and their relationships by using principal component analysis based on thin-layer chromatographic retention data. J. Chromatogr. A 822 (1998) 263-269.

- 444 [8] C. Sarbu, K. Kuhajda, S. Kevresan, Evaluation of the lipophilicity of bile acids and their
445 derivatives by thin-layer chromatography and principal component analysis. J.
446 Chromatogr. A 917 (2001) 361–366.
- 447 [9] OECD Guideline for the testing of chemicals, Test No. 107, Partition coefficient (n-
448 octanol/water), Shake flask method (1995) OECD, Paris, www.oecd.org
- 449 [10] OECD Guideline for the testing of chemicals, Test No. 117, Partition coefficient (n-
450 octanol/water), HPLC method (2004) OECD, Paris, www.oecd.org
- 451 [11] OECD Guideline for the testing of chemicals, Test No. 117, Partition coefficient (n-
452 octanol/water), Slow-stirring method (2005) OECD, Paris, www.oecd.org
- 453 [12] C. Sarbu, D. Casoni, A. Kot-Wasik, A. Wasik, J. Namiesnik, Modeling of
454 chromatographic lipophilicity of food synthetic dyes estimated on different columns. J.
455 Sep. Sci., 33(15) (2010) 2219-2229.
- 456 [13] R. D. Briciu, A. Kot-Wasik, A. Wasik, J. Namiesnik, C. Sarbu, The lipophilicity of
457 artificial and natural sweeteners estimated by reversed-phase thin-layer chromatography
458 and computed by various methods, J. Chromatogr. A, 1217 (23) (2010) 3702-3706.
- 459 [14] P. Hemstrom, K. Irgum, Hydrophilic interaction chromatography, J. Sep. Sci. 29 (2006)
460 1784-1821.
- 461 [15] M. Aleksic, J. Odovic, D. M. Milojkovic-Opsenica, Z. Lj. Tesic (2003), Salting-out
462 chromatography of several myorelaxants. J. Planar Chromatogr. Mod. TLC, 16 (2003)
463 144-146.
- 464 [16] C. Sarbu, R. D. Nascu-Briciu, D. Casoni, A. Kot-Wasik, A. Wasik, J. Namiesnik,
465 Chromatographic lipophilicity determination using large volume injections of the
466 solvents non-miscible with the mobile phase, J. Chromatogr. A, 1266 (2012) 53-60.

- [17] D. Casoni, C. Sarbu, Comprehensive evaluation of lipophilicity of biogenic amines and related compounds using different chemically bonded phases and various descriptors, *J. Sep. Sci.* 35(8) (2012) 915-921.
- [18] K. Heberger, Quantitative structure-(chromatographic) retention relationships, *J. Chromatogr. A* 1158 (2007) 273-305.
- [19] V. David, A. Medvedovici, Structure-retention correlation in liquid chromatography for pharmaceutical applications, *J. Liq. Chromatogr. Relat. Technol.* 30 (2007) 761-789.
- [20] M. Grover, B. Singh, M. Bakshi, S. Singh, Quantitative structure–property relationships in pharmaceutical research – Part 1, *Pharm. Sci. Technol. Today* 3, (2000) 28-35.
- [21] M. Grover, B. Singh, M. Bakshi, S. Singh, Quantitative structure–property relationships in pharmaceutical research – Part 2, *Pharm. Sci. Technol. Today*, 3 (2000) 50-57.
- [22] F. Tachea, R. D. Nascu-Briciu, C. Sarbu, F. Micale, A. Medvedovici, Estimation of the lipophilic character of flavonoids from the retention behavior in reversed phase liquid chromatography on different stationary phases: A comparative study, *J. Pharm. Biomed. Anal.* 57 (2012) 82-93.
- [23] V. Voicu, C. Sarbu, F. Tache, F. Micale, S. F. Radulescu, K. Sakurada, H. Ohta, A. Medvedovici, Lipophilicity indices derived from the liquid chromatographic behavior observed under bimodal retention conditions (reversed phase/hydrophilic interaction, Application to a representative set of pyridinium oximes, *Talanta*, 122 (2014) 172-179.
- [24] A. Pyka, M. Babuska, M. Zachariasz, A comparison of theoretical methods of calculation of partition coefficients for selected drugs, *Acta Pol. Pharm.* 63(3) (2006) 159-167.

- 490 [25] R. Estrada-Tejedor, N. Sabate, F. Broto, S. Nonell, I. Q. De Sarria, U. R. Llull, V.
491 Augusta, Octanol-water partition coefficients of highly hydrophobic photodynamic
492 therapy drugs : a computational study, *AFINIDAD LXX*, 564 (2013) 250-256.
- 493 [26] R. Mannhold, G. I. Poda, C. Ostermann, I. V. Tetko, Calculation of molecular
494 lipophilicity: State-of-the-art and comparison of Log P methods on more than 96,000
495 compounds, *J. Pharm. Sci.* 98(3) (2009) 861-893.
- 496 [27] D. Casoni, A. Kot-Wasik, J. Namie, C. Sarbu, Lipophilicity data for some preservatives
497 estimated by reversed-phase liquid chromatography and different computation methods,
498 *J. Chromatography A*. 1216 (2009) 2456-2465.
- 499 [28] M. Janicka, Comparison of different properties, $\log P$, $\log k_w$, and ϕ_0 as descriptors of the
500 hydrophobicity of some fungicides, *JPC – J. Planar Chromatogr. - Modern TLC*,
501 19(111) (2006) 361–370.
- 502 [29] K. Heberger, Sum of ranking differences compares methods or models fairly, *TrAC*
503 *Trends Analyt. Chem.* 29(1) (2010) 101–109.
- 504 [30] M. Vracko, N. Minovski, K. Heberger, Ranking of QSAR models to predict minimal
505 inhibitory concentrations toward *mycobacterium tuberculosis* for a set of
506 fluoroquinolones, *Acta Chim. Sloven.* 57 (2010) 586-590.
- 507 [31] K. Heberger, I. G. Zenkevich, Comparison of physicochemical and gas
508 chromatographic polarity measures for simple organic compounds, *J. Chromatogr. A*
509 1217 (2010) 2895-2902.
- 510 [32] L. Sipos, Z. Kovacs, D. Szollosi, Z. Kokai, I. Dalmadi, A. Fekete, Comparison of novel
511 sensory panel performance evaluation techniques with e-nose analysis integration, *J.*
512 *Chemom.* 25 (2011) 275-286.
- 513 [33] K. Sunjog, S. Kolarevic, K. Heberger, Z. Gacic, J. Knezevic-Vukcevic, B. Vukovic-
514 Gacic, M. Lenhardt, Comparison of comet assay parameters for estimation of

genotoxicity by sum of ranking differences, *Anal. Bioanal. Chem.* 405(14) (2013) 4879-4885.

[34] K Heberger, K. Kollar-Hunek, Sum of ranking differences for method discrimination and its validation: comparison of ranks with random numbers, *J. Chemom.* 25(4) (2011) 151-158.

[35] K. Kollar-Hunek, K. Heberger, Method and model comparison by sum of ranking differences in cases of repeated observations (ties), *Chemometr. Intell. Lab. Syst.* 127 (2013) 139-146.

[36] K. Heberger, R. Rajko, Generalization of pair correlation method (PCM) for non-parametric variable selection, *J. Chemom.* 16(8-10) (2002) 436-443.

[37] R. Rajko, K. Heberger, Conditional Fisher's exact test as a selection criterion for pair-correlation method. Type I and Type II errors. *Chemometr. Intell. Lab. Syst.* 57(1) (2001) 1-14.

529 **Tables**

530 **Table 1** Summary of chromatographically determined lipophilicity indices with short
 531 description, chromatographic technique and chromatographic modality used for its derivation.

No	Lipophilicity measure	Description	Chrom. technique / Chrom. modality	Ref.
1	$\log k$	Retention measure, so called logarithm of the retention factor (k)	- HPLC - Various modalities	[12,14,22]
2	$m\log k$	Arithmetic mean of $\log k$	- HPLC - Various modalities	[12,14,22]
3	$\log k_w$	Extrapolated $\log k$ value to the zero content of the mobile phase organic modifier ($\varphi =$ 0). Intercept in the linear equation $\log k = \log k_w + S\varphi$ but it can be also estimated from binomial dependence	- HPLC - Typical reversed-phase modality; Part of the HILIC U- shape retention profile curve that corresponds to the reversed- phase conditions	[12,14,22]

4	S	Slope in the equation $\log k = \log k_w + S\phi$ proportional to the specific hydrophobic surface of the molecule	- HPLC - Typical reversed-phase modality	[12,22]
5	ϕ_0	Concentration of the mobile phase organic modifier necessary for equal distribution of a solute between stationary and mobile phase ($\log k = 0$)	- HPLC - Typical reversed-phase modality	[12,22]
6	$PC1/\log k$	Scores corresponding to the first principal component (PCs) of $\log k$)	- HPLC - Various modalities	[12,14,22]
7	k_{\min}	Retention factor related to the minimum retention denoted on the U shaped curve of the HILIC retention profile	- HPLC - HILIC	[23]
8	$\log k_{\min}$	Logarithmic value of k_{\min}	- HPLC - HILIC	[23]
9	ISOELUT	Mobile phase composition that corresponds to the $\log k_{\min}$ value	- HPLC - HILIC	[23]
10	LOGISOELUT	Logarithmic value of	- HPLC	[23]

		ISOELUT	- HILIC	
11	HYL	Extrapolated retention property (retention factor k) on the hydrophilic interaction dominant side of the U shape HILIC retention profile curve ($\varphi = 100\%$ v/v)	- HPLC - HILIC	[23]
12	LOGHYL	Logarithmic value of HYL	- HPLC - HILIC	[23]

532

533

534 **Table 2** Summary of computationally estimated $\log P$ scales accompanied with short
 535 description.

No	$\log P$ scale	Description	Ref.
1	AlogPs	Property based, self-learning method based on the use of associative neural networks to predict the $\log P$ value from the molecular structure.	[24,25]
2	AClogP	Subgroup, atom-based method relying on 369 atom-type contribution values, obtained from 5000 molecules.	[25]
3	miLogP	Subgroup method, based on fragment contribution. It was developed using 35 small basic fragments and 185 larger fragments. Accounts for hydrogen bond contribution and charge interaction.	[24-26]
4	KOWWIN	Subgroup method; mixed both atom-based as well as fragment contribution method. Predicted $\log P$ values are obtained starting from the measured $\log P$ of structural analogues.	[24-26]
	ABlogP	Subgroup method based on fragment contributions. It applies averaged correction factors, obtained from both simple and complex compounds.	[25,26]
5	XlogP2	Subgroup, atom-based method, which uses 90 basic atom types and small number of correction factors.	[25,26]
6	XlogP3	Subgroup, atom-based approach. The main difference compared to XlogP2 method is that it starts from the known $\log P$ value of a similar reference compound.	[25,26]
7	MLOGP	Property based, Moriguchi octanol-water partition	[24,25]

coefficient - based on topological indices and
quantitative structure- $\log P$ relationships

Figure captions

Fig. 1 Loading plot of $PC1$ vs. $PC2$ of hydrophilic interaction chromatography lipophilicity indices in combination with computationally estimated $\log P$ values.

Fig. 2 Hierarchical cluster analysis dendrogram demonstrating similarities and grouping patterns of lipophilicity measures for hydrophilic interaction chromatography and computationally estimated $\log P$ values.

Fig. 3 SRD-CRRN Ranking of chromatographically estimated lipophilicity scales and computationally calculated $\log P$ values.

Fig. 4 Ranking of chromatographic lipophilicity scales and computationally estimated $\log P$ values by means of the seven fold SRD cross-validation procedure.

Fig. 5 Comparison of chromatographically derived lipophilicity indices and computationally calculated $\log P$ values by means of GPCM.

Fig. 6 Loading plot of $PC1$ vs. $PC2$ of reversed-phase HPLC lipophilicity indices and computationally estimated $\log P$ scales.

Fig. 7 Hierarchical cluster analysis dendrogram showing grouping pattern and similarities of different HPLC reversed-phase chromatographic lipophilicity indices (group A) in combination with computationally estimated $\log P$ scales (group B).

Fig. 8 SRD-CRNN ranking of chromatographically estimated lipophilicity scales and computationally calculated $\log P$ scales under reversed phase conditions.

Fig. 9 Ranking of chromatographic lipophilicity scales obtained under typical reversed-phase conditions and computationally estimated $\log P$ scales by means of the sevenfold SRD cross-validation procedure – box and whisker plot.

Fig. 10 GPCM ranking of chromatographically estimated lipophilicity scales and computationally calculated $\log P$ scales.

List of abbreviations

PC – principal component

SRD – sum of ranking (absolute) differences

CRRN – validation of the SRD procedure: Comparison of Ranks by Random Numbers.

GPCM – Generalized Pair Correlation Method. (for explanation of the abbreviations see in the text).

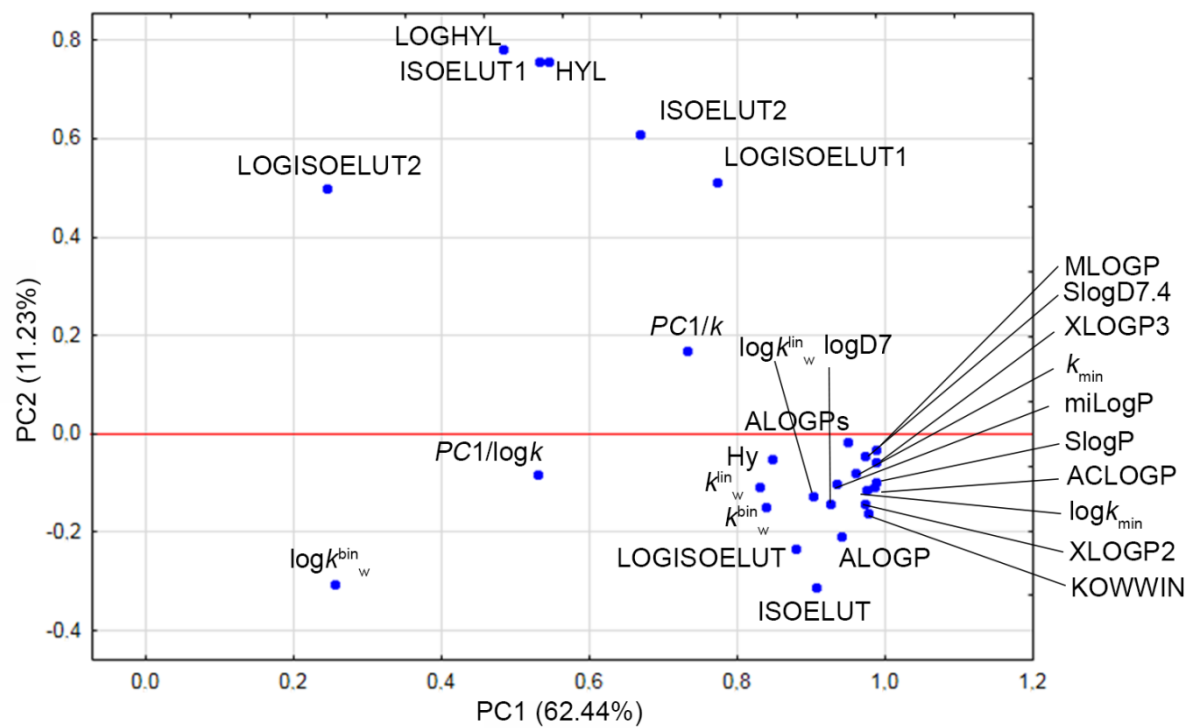


Figure 1

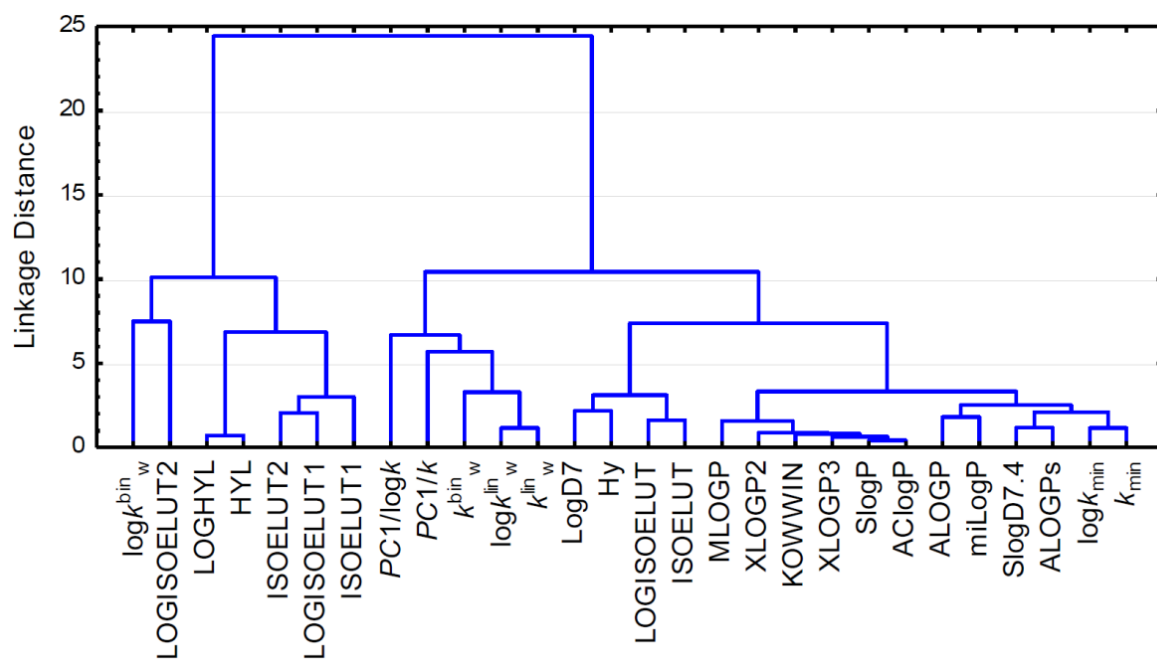


Figure 2

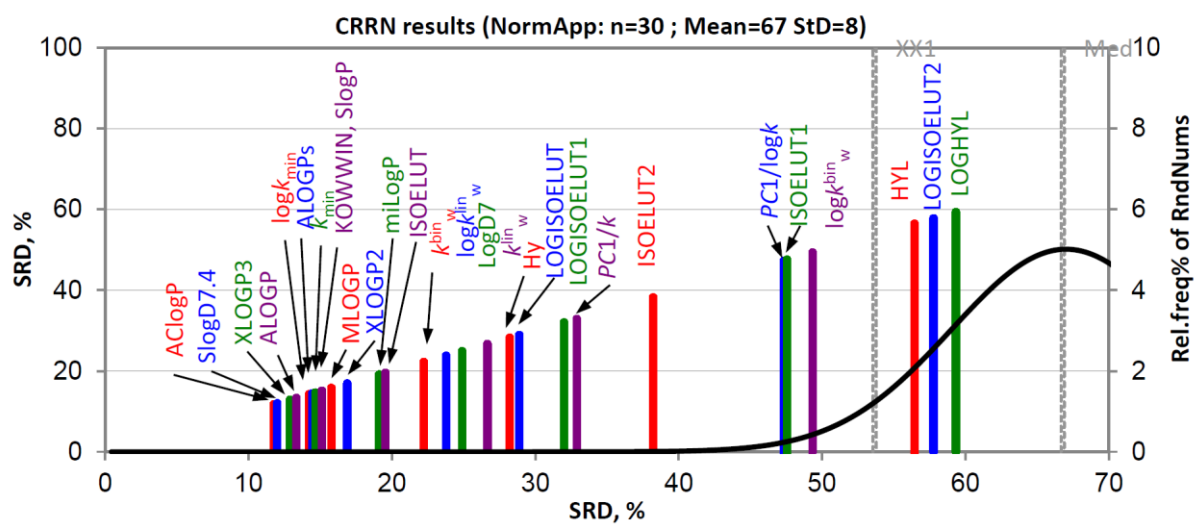


Figure 3

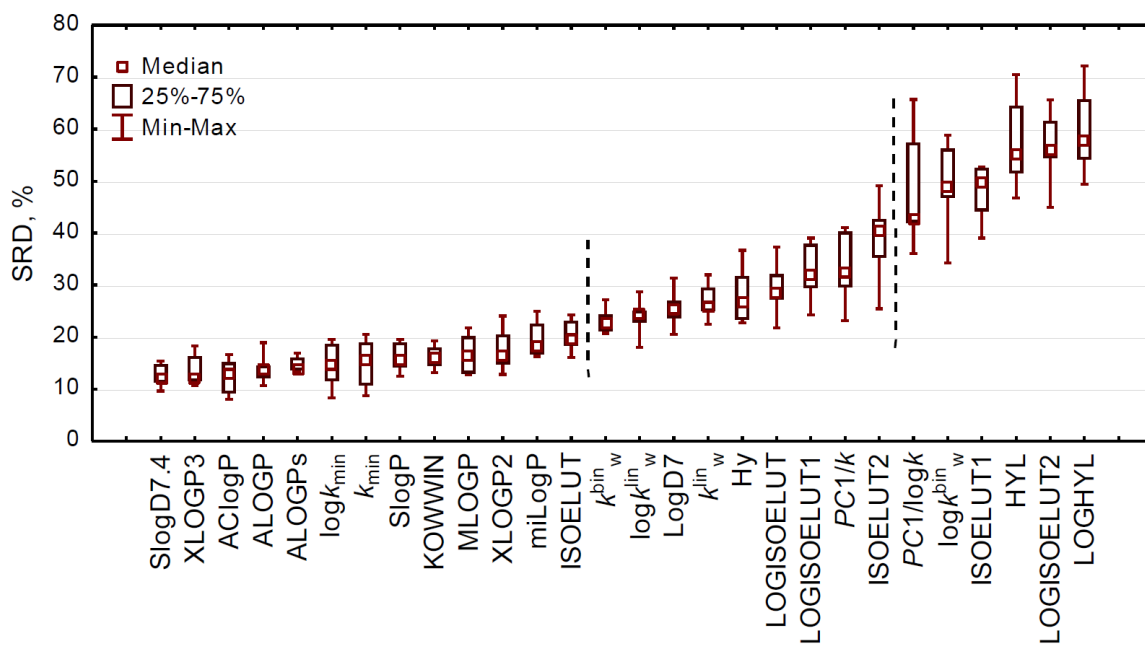


Figure 4

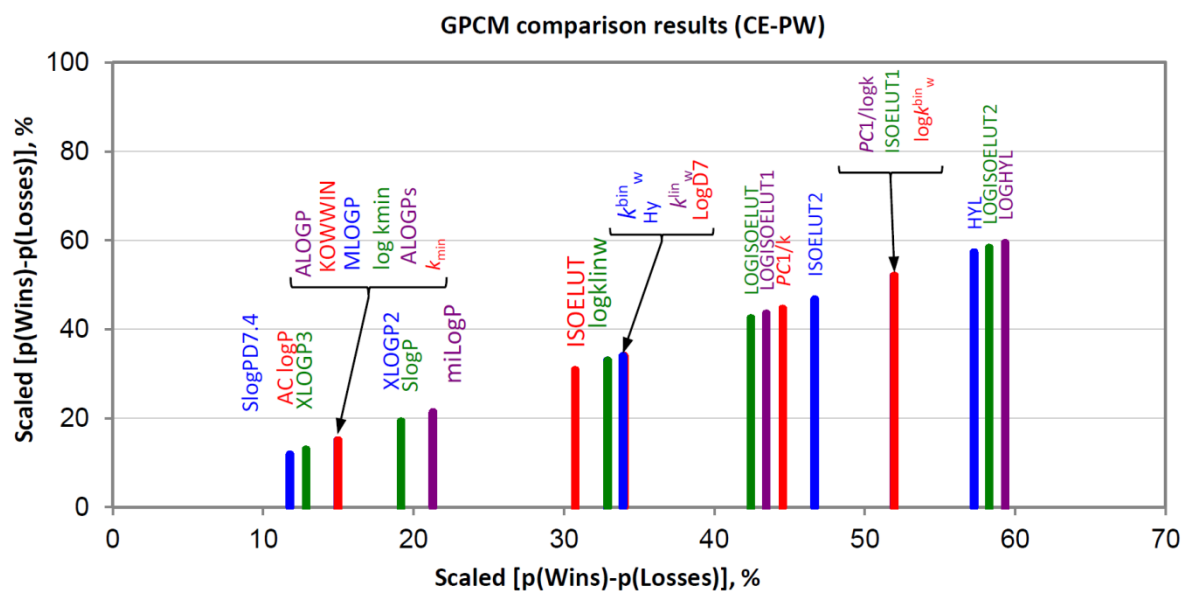


Figure 5

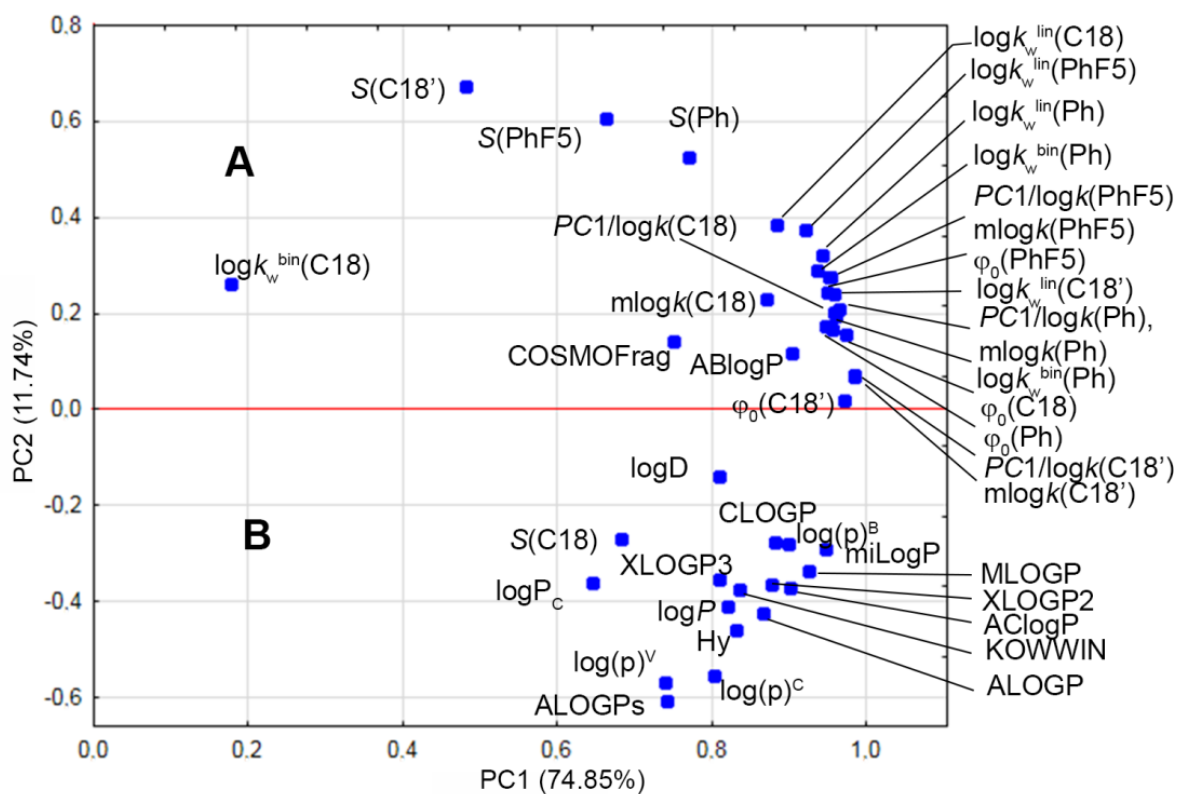


Figure 6

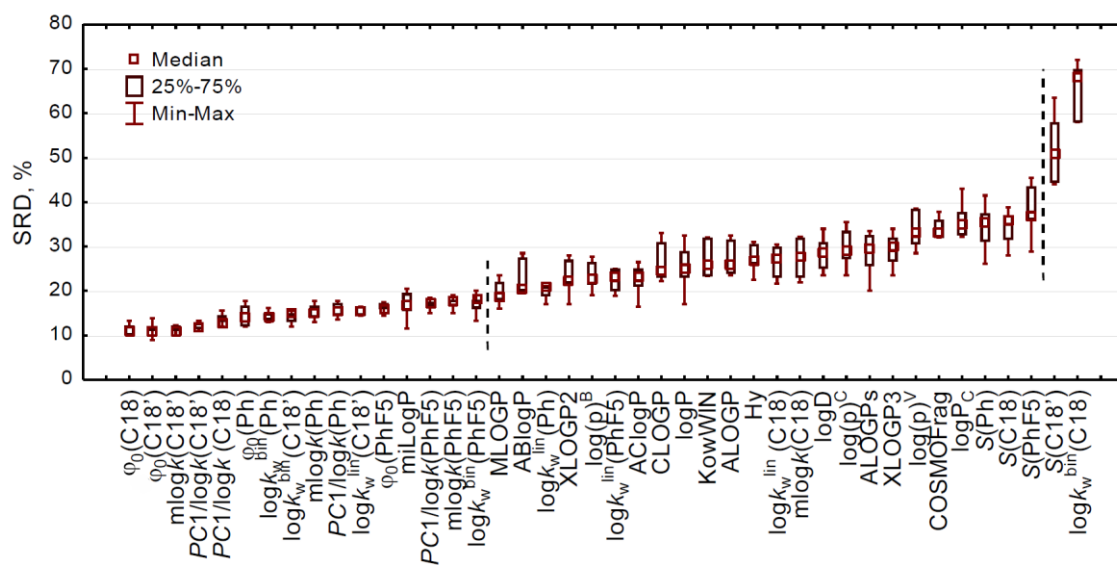


Figure 9

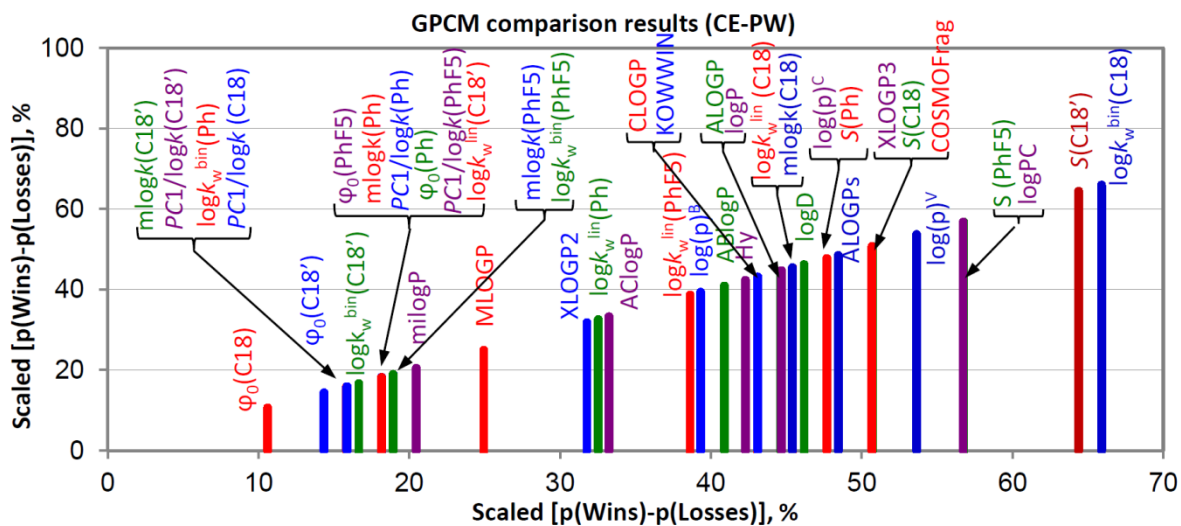


Figure 10

Supplementary material

Chromatographic and computational assessment of lipophilicity in using sum of ranking differences and generalized pair-correlation

Filip Andrić¹, Károly Héberger^{2*}

¹Faculty of Chemistry, University of Belgrade, Studentski trg 12 – 16, 11000 Belgrade, Serbia

²Research Centre for Natural Sciences, Hungarian Academy of Sciences

H-1117 Budapest XI., Magyar Tudósok krt 2., Hungary

*Corresponding author

Content

Table S1a. Case study 1 - Log P values of the target compounds obtained by various computational algorithms	p. 2
Table S1b. Case study 1 - Chromatographic lipophilicity indices of the target compounds	p. 3
Table S2. Case study 1 - Pearson's correlation coefficient values for chromatographic lipophilicity parameters (columns) and computationally estimated log P (rows)	p. 4
Table S3a. Case study 2 - Log P values of the target compounds obtained by various computational algorithms and determined chromatographic lipophilicity indices	p. 5
Table S3b. Case study 2 - Chromatographic lipophilicity indices of the target compounds obtained under different chromatographic conditions	p. 6
Table S4. Case study 2 - Values of Pearson's correlation coefficient for chromatographic lipophilicity parameters (columns) and computationally estimated log P (rows)	p. 8
Table S5. Case study 1 - Scaled rank values obtained by the SRD-CRRN and GPCM approach in the case of three different pretreatment data methods: autoscaling (AS), interval scaling (IS) and ranking (Rnk)	p. 10
Table S6. Case study 2 - Scaled rank values obtained by the SRD-CRRN and GPCM approach in the case of three different pretreatment data methods: autoscaling (AS), interval scaling (IS) and ranking (Rnk)	p. 11
Table S7. Case study 1	

Table S1a. Case study 1 - Log *P* values of the target compounds obtained by various computational algorithms

#	Comp. ^a	ALOGPs	AClogP	miLogP	KOWWIN	XLOGP2	XLOGP3	ALOGP	Hy ^b	MLOGP	SlogP	SlogD7.4	LogD7
1	2-PAE	-2.33	1.12	-2.47	-0.80	1.02	1.34	1.41	0.12	0.81	-2.37	-2.90	-4.56
2	3-PAE	-2.76	1.01	-2.76	-0.80	0.93	1.00	0.98	0.12	0.81	-2.54	-3.60	-4.06
3	4-PAE	-2.81	1.02	-3.73	-0.80	0.93	1.00	0.98	0.12	0.81	-2.40	-3.29	-4.19
4	2-PAB	-1.69	2.05	-1.41	0.18	1.95	2.22	2.39	0.20	1.43	-1.50	-2.19	-3.69
5	3-PAB	-2.22	1.94	-1.70	0.18	1.86	1.89	1.96	0.20	1.43	-1.62	-2.82	-3.20
6	2-PAH	-0.71	2.97	-0.40	1.16	3.09	3.30	3.30	0.27	2.00	-0.56	-1.37	-2.90
7	3-PAH	-1.46	2.87	-0.69	1.16	3.00	2.97	2.87	0.27	2.00	-0.66	-2.01	-2.40
8	4-PAH	-1.35	2.87	-1.65	1.16	3.00	2.97	2.87	0.27	2.00	-0.54	-1.73	-2.53
9	2-PAO	0.54	3.90	0.61	2.14	4.23	4.39	4.21	0.31	3.45	0.45	-0.47	-2.11
10	3-PAO	-0.71	3.80	0.32	2.14	4.14	4.05	3.78	0.31	3.45	0.35	-1.18	-1.61
11	4-PAO	-0.67	3.80	-0.64	2.14	4.14	4.05	3.78	0.31	3.45	0.45	-0.88	-1.74
12	2-PAD	1.51	4.83	1.62	3.13	5.36	5.47	5.12	0.35	3.95	1.47	0.46	-1.32
13	3-PAD	0.41	4.73	1.33	3.13	5.27	5.14	4.69	0.35	3.95	1.37	-0.34	-0.82
14	2-PAL	2.43	5.76	2.63	4.11	6.50	6.55	6.04	0.39	4.43	2.47	1.40	-0.52
15	3-PAL	1.40	5.66	2.35	4.11	6.41	6.22	5.61	0.39	4.43	2.37	0.50	-0.03
16	4-PAL	1.27	5.66	1.38	4.11	6.41	6.22	5.61	0.39	4.43	2.44	0.87	-0.15
17	2-PABn	-1.10	2.34	-1.25	0.41	1.96	2.63	2.64	0.29	2.02	-1.08	-1.69	-3.13
18	3-PABn	-1.79	2.23	-1.54	0.41	1.87	2.30	2.21	0.29	2.02	-1.22	-2.35	-2.63
19	4-PABn	-2.00	2.23	-2.51	0.41	1.88	2.75	0.32	0.29	1.95	-1.07	-2.07	-2.75
20	2-PAPE	-0.74	2.40	-1.04	0.91	2.47	2.93	2.96	0.31	2.29	-0.84	-1.54	-2.87
21	3-PAPE	-1.81	2.29	-1.33	0.91	2.39	2.59	2.53	0.31	2.29	-0.97	-2.19	-2.38
22	4-PAPE	-1.82	2.29	-2.30	0.91	2.38	2.59	2.53	0.31	2.29	-0.82	-1.89	-2.50
23	3-PAPP	-1.59	2.76	-0.81	1.40	2.74	2.95	2.99	0.33	2.54	-0.60	-1.88	-1.98
24	3-PAPB	-1.31	3.22	-0.54	1.89	3.31	3.31	3.45	0.35	2.79	-0.16	-1.50	-1.58
25	4-PAPB	-1.34	3.22	-1.51	1.89	3.31	3.31	3.45	0.35	2.79	-0.02	-1.23	-1.71
26	2-PAMB	-0.78	2.65	-0.81	0.96	2.40	3.00	3.13	0.31	2.29	-0.71	-1.34	-2.66
27	3-PAMB	-1.48	2.55	-1.09	0.96	2.31	2.66	2.70	0.31	2.29	-0.86	-2.00	-2.16
28	4-PAMB	-1.71	2.55	-2.06	0.96	2.31	2.66	2.70	0.31	2.29	-0.71	-1.71	-2.29
29	3-PATB	-0.35	3.75	0.17	2.32	3.69	3.97	3.61	0.37	3.04	0.41	-0.92	-1.00
30	4-PATB	-0.22	3.75	-0.80	2.32	3.69	3.97	3.61	0.37	3.04	0.55	-0.62	-1.13

^a Derivatives of mono-pyridinium oxime compounds and their abbreviations are given in Figure 1 of reference [1] (ref. [26] in the manuscript).^b Variables multiplied by -1

Table S1b. Case study 1 - Chromatographic lipophilicity indices of the target compounds

#	Comp. ^a	k _{min}	log k _{min}	ISOELUT	LOGISOELUT	ISOELUT1	LOGISOELUT1	ISOELUT2	LOGISOELUT2	k _w ^{lin}	logk _w ^{lin}	k _w ^{bin}	logk _w ^{bin}	HYL	LOGHYL	PC1/k	PC1/logk
1	2-PAE	2.82	0.45	37.22	33.33	52.96	42.19	56.56	51.04	6.97	0.88	8.85	1.00	18.75	1.41	-17.52	0.48
2	3-PAE	2.66	0.43	36.18	32.33	55.63	43.53	56.77	52.31	6.08	0.82	6.91	0.86	19.45	1.43	-17.24	0.44
3	4-PAE	2.81	0.45	37.47	34.00	54.76	44.29	58.11	96.98	6.82	0.88	7.87	0.92	18.79	1.41	-17.78	0.48
4	2-PAB	4.01	0.60	48.15	48.83	38.18	40.35	57.80	69.46	18.26	1.35	24.57	1.48	16.26	1.32	-17.93	0.71
5	3-PAB	3.71	0.57	46.62	45.33	41.53	41.53	58.53	59.18	14.29	1.23	19.07	1.36	15.65	1.31	-19.20	0.65
6	2-PAH	5.59	0.75	58.51	62.33	40.84	49.67	62.41	66.78	38.97	1.74	52.10	1.76	15.05	1.24	-14.58	0.93
7	3-PAH	4.80	0.68	56.52	58.33	44.79	52.76	66.18	72.87	29.22	1.59	41.91	1.73	15.40	1.28	-16.28	0.86
8	4-PAH	5.17	0.71	57.58	59.17	41.58	49.51	66.93	66.81	32.83	1.65	47.94	1.81	14.44	1.23	-16.05	0.89
9	2-PAO	6.03	0.78	61.56	58.38	64.91	66.77	72.79	74.72	82.13	2.52	75.97	1.18	48.67	1.86	18.48	1.16
10	3-PAO	5.64	0.75	60.83	60.71	66.43	66.81	71.86	69.95	64.44	2.36	126.84	2.50	44.69	1.82	13.70	1.09
11	4-PAO	5.57	0.75	61.32	64.43	65.19	66.83	72.81	69.19	71.76	2.43	102.47	1.62	44.34	1.82	13.91	1.11
12	2-PAD	7.46	0.87	66.18	66.89	69.76	67.48	69.47	69.15	86.07	3.06	77.91	0.11	45.89	1.80	17.27	1.13
13	3-PAD	6.90	0.84	66.43	70.75	70.03	68.87	72.07	70.55	73.71	2.77	61.19	0.99	40.03	1.74	11.85	1.09
14	2-PAL	9.66	0.98	70.50	69.35	78.03	76.69	78.13	78.75	57.35	2.34	82.94	1.58	45.59	1.76	23.15	1.37
15	3-PAL	8.85	0.95	69.86	70.40	77.88	76.23	77.08	77.10	50.38	2.27	91.93	2.22	44.17	1.76	19.45	1.32
16	4-PAL	8.70	0.94	70.21	70.80	78.16	76.77	78.37	75.96	50.88	2.27	80.39	1.70	43.49	1.76	18.77	1.32
17	2-PABn	4.25	0.63	46.25	49.63	80.00	72.83	79.56	80.03	15.88	1.26	20.96	1.40	55.31	1.95	-6.29	0.23
18	3-PABn	4.06	0.61	45.25	47.38	79.36	70.01	78.56	74.61	14.16	1.23	17.51	1.26	51.34	1.91	-2.24	-0.13
19	4-PABn	4.15	0.62	45.95	47.88	78.89	70.07	77.88	80.16	14.93	1.26	17.83	1.26	49.29	1.88	-2.06	0.05
20	2-PAPE	4.63	0.67	55.35	59.17	54.68	57.37	70.00	70.64	24.02	1.51	27.46	1.39	18.72	1.40	-6.04	0.22
21	3-PAPE	4.45	0.65	54.74	57.83	55.04	56.81	69.75	79.82	21.46	1.46	22.39	1.27	17.47	1.36	-7.05	0.38
22	4-PAPE	4.48	0.65	55.77	60.67	54.39	57.66	71.67	72.88	23.76	1.51	25.80	1.34	17.67	1.37	28.39	-0.77
23	2-PAMB	4.93	0.69	57.60	52.75	54.00	58.59	67.97	68.09	26.28	1.56	40.79	1.97	16.20	1.31	30.13	-0.80
24	3-PAMB	4.73	0.68	57.03	52.63	44.16	49.69	60.05	63.73	31.98	1.77	31.96	1.36	15.23	1.27	-3.18	-0.11
25	4-PAMB	4.71	0.67	58.14	51.38	42.61	51.90	64.95	64.47	36.26	1.73	43.47	1.62	14.85	1.26	-2.98	-0.13
26	3-PAPP	4.79	0.68	54.11	50.80	79.10	72.21	77.81	75.64	22.24	1.53	26.58	1.32	41.72	1.70	-6.29	0.23
27	3-PAPB	5.04	0.70	60.22	57.63	66.58	66.53	67.44	71.04	15.61	1.32	26.42	1.71	14.55	1.25	-6.29	0.23
28	4-PAPB	5.04	0.70	60.65	59.13	66.07	66.21	66.89	68.82	16.43	1.35	29.41	1.83	14.34	1.24	-4.80	0.19
29	3-PATB	5.39	0.73	65.46	70.88	68.12	68.47	71.41	70.62	29.46	1.75	43.46	1.85	16.92	1.30	-5.32	0.17
30	4-PATB	5.56	0.75	67.72	65.70	62.87	66.28	69.33	70.78	42.78	2.02	71.65	2.06	14.71	1.24	-5.98	0.22

^a Derivatives of mono-pyridinium oxime compounds and their abbreviations are given in Figure 1 of reference [1] (ref. [26] in the manuscript).

Table S2. Case study 1 - Values of Pearson's correlation coefficient for chromatographic lipophilicity parameters (columns) and computationally estimated log*P* (rows)

	<i>k</i> _{min}	log <i>k</i> _{min}	ISOELUT	LOGISOELUT	ISOELUT1	LOGISOELUT1	ISOELUT2	LOGISOELUT2	<i>k</i> _w ^{lin}	log <i>k</i> _w ^{lin}	<i>k</i> _w ^{bin}	log <i>k</i> _w ^{bin}	HYL	LOGHYL	PC1/ <i>k</i>	PC1/log <i>k</i>
ALOGPs	0.9536	0.9336	0.8305	0.7960	0.4940	0.6844	0.5740	0.2027	0.7954	0.8558	0.7490	0.1012	0.5247	0.4624	0.6268	0.6076
AC log <i>P</i>	0.9703	0.9701	0.9051	0.8602	0.4462	0.6884	0.5704	0.1824	0.8125	0.8881	0.8243	0.2606	0.4680	0.4030	0.6747	0.5660
miLog <i>P</i>	0.9293	0.9222	0.8280	0.7949	0.4183	0.6273	0.5099	0.0996	0.7863	0.8545	0.7643	0.1866	0.4797	0.4214	0.6170	0.6056
KOWWIN	0.9551	0.9651	0.9318	0.8787	0.4088	0.6745	0.5555	0.1719	0.7918	0.8776	0.8074	0.2913	0.3963	0.3319	0.6949	0.4857
XLOGP2	0.9692	0.9568	0.8852	0.8405	0.3966	0.6313	0.5186	0.1642	0.8289	0.8942	0.8347	0.2303	0.4469	0.3913	0.6803	0.6074
XLOGP3	0.9771	0.9723	0.8900	0.8545	0.4793	0.7089	0.6020	0.2062	0.8113	0.8855	0.8090	0.2192	0.5048	0.4429	0.6921	0.5687
ALOG <i>P</i>	0.9356	0.9362	0.8911	0.8314	0.3357	0.5922	0.4721	0.1116	0.7980	0.8662	0.7882	0.2314	0.3696	0.3059	0.6456	0.5472
Hy	0.7695	0.8563	0.8915	0.8637	0.4425	0.7598	0.6974	0.2392	0.5760	0.6998	0.5926	0.3788	0.3142	0.2341	0.6526	0.0759
MLOG <i>P</i>	0.9290	0.9486	0.9075	0.8619	0.5053	0.7504	0.6348	0.2100	0.8310	0.9030	0.8293	0.2493	0.5070	0.4530	0.7691	0.4584
Slog <i>P</i>	0.9684	0.9710	0.9158	0.8717	0.4597	0.7067	0.5929	0.2039	0.7977	0.8795	0.8113	0.2647	0.4558	0.3917	0.6960	0.5218
SlogD7.4	0.9627	0.9622	0.8882	0.8448	0.4841	0.7174	0.6115	0.2208	0.7901	0.8648	0.7729	0.1905	0.4913	0.4217	0.6731	0.5237
LogD7	0.8789	0.9127	0.9281	0.8768	0.4399	0.7153	0.5908	0.1896	0.6830	0.7926	0.7352	0.3819	0.3294	0.2620	0.6734	0.3164

Statistically significantly correlated variables ($p = 0.05$) are marked in bold.

Table S3a. Case study 2 - LogP values of the target compounds obtained by various computational algorithms and determined chromatographic lipophilicity indices

Comp a	logDa	log(p)C	log(p)V	log(p)B	CLOGP	logPC	logP	Hy	MLOGP	ALOGP	ALOGPs	AClogP	ABlog P	COSMOF raq	mil gP	KowWIN	XLOGP 2	XLOGP 3
1	2.97	3.07	3.17	3.04	3.48	2.75	2.21	0.86	3.15	3.14	3.10	3.59	3.42	3.11	3.74	3.51	3.21	3.56
2	2.81	2.95	2.92	3.17	2.89	0.96	1.87	0.83	2.83	3.12	3.14	3.49	3.38	3.81	3.75	3.59	3.12	3.53
3	2.83	2.27	2.20	3.39	3.09	1.69	1.93	0.83	2.83	2.64	2.83	3.45	3.70	3.80	3.72	2.80	2.96	3.73
4	2.81	2.95	2.92	3.17	3.59	1.03	1.90	0.83	2.83	3.12	3.11	3.49	3.41	2.90	3.75	3.59	3.12	3.47
5	2.81	2.95	2.92	3.17	3.59	1.27	2.01	0.83	2.83	3.12	3.18	3.49	3.25	3.08	3.77	3.59	3.12	3.95
6	2.81	2.95	2.92	3.17	3.59	1.07	2.36	0.83	2.83	3.12	3.18	3.49	3.43	3.43	3.77	3.59	3.12	3.95
7	2.65	2.82	2.66	3.30	3.27	0.43	2.80	0.80	2.53	3.10	3.09	3.38	3.06	3.67	3.56	3.67	2.77	3.12
8	2.56	1.78	1.67	3.05	2.87	0.36	1.90	0.32	2.29	2.56	2.47	2.89	3.80	4.18	3.48	2.70	2.79	3.37
9	3.31	2.68	2.88	2.65	4.11	2.52	1.67	0.33	3.09	2.87	3.37	3.29	3.46	3.56	3.47	3.80	2.80	2.45
10	2.66	2.68	2.88	2.65	3.21	2.55	1.68	0.33	2.58	2.87	3.31	3.29	2.89	1.97	3.23	3.03	2.80	3.62
11	2.66	2.68	2.88	2.65	3.21	2.47	1.69	0.33	2.58	2.87	3.27	3.29	3.14	2.52	3.23	3.03	2.80	3.62
12	2.42	1.52	1.64	2.53	2.30	0.37	1.19	-0.34	2.04	2.31	2.29	2.70	3.27	2.87	2.94	2.14	2.47	3.05
13	2.42	1.52	1.64	2.53	2.30	0.35	1.17	-0.34	2.04	2.31	2.24	2.70	3.51	3.29	2.94	2.14	2.47	3.05
14	2.71	1.90	2.32	1.87	2.91	0.31	0.75	-1.10	1.76	2.33	2.47	2.69	2.43	2.44	2.46	2.84	1.15	1.74
15	2.71	1.90	2.32	1.87	3.00	0.31	1.19	-1.10	1.76	2.33	2.66	2.69	2.84	3.02	2.68	3.27	2.64	1.74
16	2.76	1.13	1.35	2.14	2.56	0.31	1.05	-1.10	1.76	2.04	1.46	2.40	3.42	2.73	2.65	2.44	2.06	2.25
17	2.46	0.74	1.07	1.75	1.90	0.31	0.94	-1.92	0.99	1.77	1.23	2.10	2.80	2.03	2.17	1.96	0.81	1.90
18	2.40	1.51	2.03	1.48	2.31	0.31	0.36	-1.92	0.99	2.07	2.15	2.40	1.96	2.36	1.97	2.36	0.75	1.38
19	2.16	0.35	0.78	1.36	1.30	0.57	0.30	-2.78	0.23	1.50	1.07	1.80	2.34	1.94	1.68	1.48	0.41	1.54
20	1.96	1.00	1.10	2.27	1.49	0.39	0.50	-1.07	0.73	2.02	1.73	2.30	2.74	3.37	2.28	1.49	1.56	2.30
21	2.73	2.13	2.25	2.20	2.08	0.33	1.87	-0.34	1.77	2.33	2.77	2.24	2.52	1.90	2.56	2.55	2.08	2.47
22	3.48	3.56	3.64	3.45	3.98	2.80	2.38	0.87	3.40	3.62	3.53	3.91	3.83	3.41	4.16	4.06	3.65	3.92
23	4.08	4.12	4.15	4.07	4.69	3.15	2.66	0.83	3.91	4.29	3.92	4.52	4.54	3.99	4.77	4.70	4.06	4.55

^a Derivatives of flavonoids and their identification numbers are given in Figure 1 of reference [2] (ref. [25] in the manuscript).

Table S3b. Case study 2 – Chromatographic lipiphilicity indices of the target compounds obtained under different chromatographic conditions

Comp. ^a	Column C18						Column C18'					
	$\log k_w^{\text{lin}}$ (C18)	$\log k_w^{\text{bin}}$ (C18)	mlogk(C18)	S(C18) [#]	$\varphi 0$ (C18) [#]	PC1/logk (C18) [#]	$\log k_w^{\text{lin}}$ (C18')	$\log k_w^{\text{bin}}$ (C18')	mlogk(C18')	S(C18') [#]	$\varphi 0$ (C18') [#]	PC1/logk(C18') [#]
1	2.62	3.65	0.857	0.044	59.5	0.53	2.14	2.85	0.575	0.035	61.5	0.44
2	2.67	2.64	0.688	0.044	60.7	0.66	2.28	3.08	0.628	0.037	62.1	0.56
3	2.76	2.58	0.778	0.044	62.7	0.84	2.24	2.86	0.594	0.036	61.3	0.48
4	2.62	3.92	0.758	0.050	52.8	0.04	1.93	2.65	0.445	0.033	58.5	0.15
5	2.76	2.84	0.742	0.045	61.5	0.77	2.36	3.13	0.670	0.038	62.8	0.65
6	2.66	2.74	0.669	0.044	60.1	0.62	2.37	3.15	0.691	0.037	63.5	0.70
7	2.67	3.78	0.802	0.047	57.2	0.41	2.26	3.10	0.627	0.036	62.3	0.56
8	3.05	2.89	0.869	0.048	62.9	1.04	2.49	3.24	0.676	0.040	61.8	0.67
9	2.99	3.93	0.798	0.044	68.2	2.15*	2.54	3.16	0.758	0.040	64.1	0.85
10	2.49	3.35	0.625	0.050	50	-0.21	1.80	2.32	0.248	0.035	52.2	-0.29
11	2.55	3.63	0.658	0.054	47.2	-0.46	1.86	2.61	0.321	0.034	54.4	-0.13
12	2.80	4.00	0.801	0.053	52.6	0.1	1.96	2.62	0.289	0.037	52.8	-0.20
13	2.69	3.74	0.735	0.052	51.6	-0.02	1.95	2.45	0.288	0.037	52.8	-0.20
14	2.51	3.60	0.552	0.056	44.9	-0.7	1.79	2.54	0.107	0.037	47.9	-0.6
15	2.52	3.62	0.634	0.054	46.8	-0.5	1.61	2.25	0.095	0.034	47.8	-0.63
16	2.90	3.92	0.863	0.051	56.9	0.51	2.27	2.95	0.462	0.040	56.5	0.19
17	2.55	3.59	0.603	0.056	45.8	-0.59	1.80	2.18	0.106	0.038	47.8	-0.61
18	2.14	3.23	0.248	0.054	39.6	-1.28	1.50	1.75	-0.147	0.037	41.0	-1.17
19	2.10	3.13	0.267	0.052	40.1	-1.21	1.50	1.43	-0.161	0.037	40.6	-1.21
20	2.10	3.27	0.321	0.051	41.3	-1.09	1.60	1.47	-0.063	0.037	43.3	-0.98
21	1.78	2.61	0.140	0.047	38	-1.38	1.34	1.05	-0.213	0.034	38.8	-1.32
22	2.93	3.02	0.866	0.046	63.9	1.02	2.46	3.23	0.736	0.038	64.2	0.80
23	3.33	4.55	0.972	0.047	70.6	2.66*	2.92	3.84	0.955	0.044	66.9	1.29

^a Derivatives of flavonoids and their identification numbers are given in Figure 1 of reference [2] (ref. [25] in the manuscript).

*Missing value replaced by the estimated one according to appropriate retention on C18' stationary phase

[#]Variables multiplied by -1

Table S3b. Continues

Comp. ^a	Column Ph						Column PhF5					
	$\log k_w^{\text{lin}}(\text{Ph})$	$\log k_w^{\text{bin}}(\text{Ph})$	$\text{mlog}k(\text{Ph})$	$S(\text{Ph})^{\#}$	$\varphi 0(\text{Ph})^{\#}$	$\text{PC1}/\log k(\text{Ph})^{\#}$	$\log k_w^{\text{lin}}(\text{PhF5})$	$\log k_w^{\text{bin}}(\text{PhF5})$	$\text{mlog}k(\text{PhF5})$	$S(\text{PhF5})^{\#}$	$\varphi 0(\text{PhF5})^{\#}$	$\text{PC1}/\log k(\text{PhF5})^{\#}$
1	2.00	2.51	0.505	0.865	3.55	2.82	2.35	2.98	0.624	0.43	3.40	2.62
2	2.17	2.82	0.577	1.733	3.80	3.23	2.52	3.22	0.689	1.73	3.54	2.90
3	2.23	2.65	0.647	1.733	4.27	3.62	2.64	2.97	0.773	1.73	3.85	3.26
4	1.81	2.46	0.340	0.865	2.47	1.87	2.06	2.76	0.372	0.00	2.17	1.58
5	2.22	2.82	0.587	2.167	3.78	3.28	2.60	3.20	0.721	2.17	3.60	3.03
6	2.15	2.73	0.540	2.167	3.53	3.03	2.54	3.12	0.683	1.73	3.44	2.88
7	2.06	2.62	0.514	1.299	3.51	2.87	2.39	2.95	0.564	1.73	2.94	2.38
8	2.47	3.27	0.698	3.468	4.14	3.92	2.85	3.55	0.907	2.60	4.31	3.82
9	2.49	3.01	0.783	3.035	4.77	4.41	3.03	3.58	1.035	3.03	4.77	4.34
10	1.75	2.19	0.229	1.299	1.66	1.23	1.98	2.56	0.246	0.87	1.50	1.05
11	1.63	2.04	0.156	0.865	1.21	0.82	1.91	2.47	0.188	0.43	1.21	0.81
12	1.99	2.45	0.370	2.167	2.45	2.05	2.29	2.85	0.445	1.73	2.36	1.88
13	1.90	2.28	0.320	1.733	2.18	1.74	2.23	2.63	0.433	1.30	2.34	1.84
14	1.62	1.67	0.091	1.299	0.73	0.44	2.02	2.45	0.169	1.73	1.05	0.74
15	1.60	1.84	0.147	0.432	1.16	0.75	1.93	1.90	0.177	0.87	1.15	0.77
16	2.24	2.61	0.522	3.035	3.21	2.92	2.71	3.35	0.684	3.47	3.19	2.88
17	1.72	1.90	0.143	1.733	1.05	0.75	2.10	2.44	0.224	2.17	1.29	0.96
18	1.36	1.46	-0.129	0.865	-0.76	-0.82	1.67	2.06	-0.097	0.87	-0.22	-0.38
19	1.40	1.45	-0.084	0.865	-0.44	-0.56	1.73	1.65	-0.059	1.30	-0.02	-0.22
20	1.27	1.09	-0.050	-0.870	-0.26	-0.38	1.57	1.69	-0.017	-0.87	0.17	-0.05
21	1.23	1.19	-0.135	-0.436	-0.87	-0.87	1.43	1.37	-0.214	-0.44	-0.87	-0.87
22	2.30	3.10	0.641	2.601	4.05	3.59	2.73	3.33	0.817	2.60	3.96	3.44
23	2.73	3.54	0.845	4.770	4.68	4.77	3.31	4.08	1.134	4.77	4.77	4.77

^a Derivatives of flavonoids and their identification numbers are given in Figure 1 of reference [2] (ref. [25] in the manuscript).

[#]Variables multiplied by -1

Table S4. Case study 2 - Values of Pearson's correlation coefficient for chromatographic lipophilicity parameters (columns) and computationally estimated $\log P$ (rows)

	$\log k_w^{\text{lin}}(\text{C18})$	$\log k_w^{\text{bin}}(\text{C18})$	$\text{mlog}k(\text{C18})$	$S(\text{C18})$	$\text{fi}0(\text{C18})$	$\text{PC1}/\log k(\text{C18})$	$\log k_w^{\text{lin}}(\text{C18}')$	$\log k_w^{\text{bin}}(\text{C18}')$	$\text{mlog}k(\text{C18}')$	$S(\text{C18}')$	$\text{fi}0(\text{C18}')$	$\text{PC1}/\log k(\text{C18}')$
$\log Da$	0.6644	0.2878	0.5962	0.5028	0.7362	0.7979	0.7328	0.7004	0.7232	0.4667	0.6811	0.7230
$\log(p)C$	0.5007	0.0708	0.5395	0.6382	0.6745	0.6611	0.6408	0.6768	0.7371	0.0855	0.7422	0.7372
$\log(p)V$	0.4545	0.1199	0.4774	0.5461	0.6026	0.6066	0.5702	0.6237	0.6603	0.0736	0.6619	0.6604
$\log(p)B$	0.6706	0.0220	0.7151	0.7293	0.8288	0.7865	0.8088	0.7803	0.8745	0.2518	0.8764	0.8745
CLOGP	0.7117	0.2543	0.7053	0.5454	0.7939	0.8066	0.7673	0.8180	0.8314	0.2656	0.8258	0.8316
$\log PC$	0.4494	0.1812	0.4595	0.4803	0.5713	0.5997	0.5246	0.4960	0.5760	0.1795	0.5707	0.5749
$\log P$	0.5335	-0.0230	0.6212	0.7318	0.7275	0.6805	0.6965	0.6927	0.7918	0.0899	0.7999	0.7917
Hy	0.5500	-0.0711	0.6513	0.7481	0.7499	0.6808	0.6854	0.7074	0.8084	0.0041	0.8351	0.8085
MLOGP	0.7080	0.1328	0.7539	0.6766	0.8406	0.8194	0.7882	0.8187	0.8742	0.1948	0.8790	0.8741
ALOGP	0.6123	0.1545	0.6267	0.6223	0.7461	0.7452	0.7396	0.7525	0.8057	0.2377	0.7988	0.8058
ALOGPs	0.4428	0.0363	0.4730	0.5994	0.6132	0.6092	0.5575	0.6037	0.6655	0.0000	0.6729	0.6657
AClogP	0.6765	0.1433	0.7035	0.6296	0.8014	0.7821	0.7739	0.8095	0.8533	0.2171	0.8574	0.8533
ABlogP	0.8639	0.2282	0.8573	0.5578	0.8860	0.8859	0.8797	0.8276	0.8816	0.4943	0.8585	0.8813
COSMOFraq	0.6845	0.0065	0.6481	0.5971	0.7847	0.7498	0.7758	0.7233	0.7739	0.4401	0.7597	0.7751
$\text{milog}P$	0.7303	0.0812	0.7653	0.7138	0.8707	0.8390	0.8375	0.8398	0.9109	0.2588	0.9117	0.9108
KowWIN	0.6047	0.1685	0.6123	0.5905	0.7354	0.7376	0.7086	0.7592	0.7839	0.1991	0.7799	0.7842
XLOGP2	0.6608	0.0833	0.7348	0.6763	0.7935	0.7594	0.7300	0.7480	0.8285	0.1205	0.8355	0.8285
XLOGP3	0.5840	-0.0538	0.6453	0.6501	0.7209	0.6664	0.7011	0.6915	0.7822	0.1484	0.7923	0.7815
	$\log k_w^{\text{lin}}(\text{Ph})$	$\log k_w^{\text{bin}}(\text{Ph})$	$\text{mlog}k(\text{Ph})$	$S(\text{Ph})$	$\text{fi}0(\text{Ph})$	$\text{PC1}/\log k(\text{Ph})$	$\log k_w^{\text{lin}}(\text{PhF5})$	$\log k_w^{\text{bin}}(\text{PhF5})$	$\text{mlog}k(\text{PhF5})$	$S(\text{PhF5})$	$\text{fi}0(\text{PhF5})$	$\text{PC1}/\log k(\text{PhF5})$
$\log Da$	0.7094	0.7118	0.6965	0.6620	0.6616	0.6962	0.7277	0.7018	0.7066	0.6400	0.6668	0.7072
$\log(p)C$	0.5705	0.6544	0.6300	0.3723	0.6358	0.6298	0.5360	0.6040	0.5941	0.2638	0.5980	0.5941
$\log(p)V$	0.5017	0.5836	0.5489	0.3467	0.5518	0.5488	0.4793	0.5454	0.5224	0.2607	0.5214	0.5225
$\log(p)B$	0.7545	0.8100	0.8163	0.5125	0.8192	0.8158	0.7063	0.7491	0.7776	0.3662	0.7851	0.7777
CLOGP	0.7298	0.7760	0.7593	0.5864	0.7572	0.7592	0.7184	0.7564	0.7466	0.4949	0.7425	0.7467

logPC	0.4756	0.5135	0.5092	0.3671	0.5043	0.5092	0.4731	0.5016	0.5093	0.2763	0.5022	0.5091
logP	0.6530	0.7265	0.7194	0.4115	0.7312	0.7192	0.5986	0.6296	0.6631	0.3044	0.6738	0.6632
Hy	0.6446	0.7216	0.7382	0.3443	0.7643	0.7377	0.5783	0.6543	0.6799	0.1718	0.7096	0.6794
MLOGP	0.7648	0.8224	0.8174	0.5546	0.8257	0.8171	0.7262	0.7779	0.7843	0.4168	0.7935	0.7842
ALOGP	0.6638	0.7383	0.7066	0.4910	0.7009	0.7065	0.6340	0.6953	0.6816	0.3779	0.6768	0.6818
ALOGPs	0.5043	0.5884	0.5696	0.3019	0.5765	0.5690	0.4629	0.5230	0.5294	0.1756	0.5354	0.5294
AClogP	0.7128	0.7802	0.7646	0.5143	0.7710	0.7645	0.6827	0.7436	0.7384	0.3939	0.7453	0.7385
ABlogP	0.8799	0.8879	0.8915	0.7393	0.8732	0.8911	0.8579	0.8452	0.8886	0.6029	0.8781	0.8890
COSMOFraq	0.7422	0.7270	0.7876	0.5121	0.7778	0.7863	0.7310	0.6950	0.7873	0.4240	0.7903	0.7875
milogP	0.7968	0.8583	0.8504	0.5744	0.8537	0.8500	0.7553	0.7975	0.8178	0.4319	0.8250	0.8179
KowWIN	0.6591	0.7218	0.6988	0.4919	0.6992	0.6986	0.6434	0.6774	0.6742	0.4356	0.6710	0.6744
XLOGP2	0.7097	0.7823	0.7791	0.4586	0.7884	0.7783	0.6557	0.6900	0.7350	0.3026	0.7479	0.7351
XLOGP3	0.6526	0.7331	0.7057	0.4546	0.7145	0.7057	0.5898	0.6625	0.6642	0.2682	0.6794	0.6642

Statistically significantly correlated variables ($p = 0.05$) are marked in bold.

Table S6. Case study 2 – Scaled rank values obtained by the SRD-CRRN and GPCM approach in the case of three different pretreatment data methods: autoscaling (AS), interval scaling (IS) and ranking (Rnk).

SRD scores				PCM scores (RScale)							
AS		IS		Rnk		AS		IS		Rnk	
Variable		Variable		Variable		Variable		Variable		Variable	
φ0(C18)	10,61	φ0(C18)	10,61	φ0(C18)	9,09	φ0(C18)	10,61	φ0(C18)	10,61	φ0(C18)	9,09
mlogk(C18')	10,61	mlogk(C18')	10,61	φ0(C18')	9,09	φ0(C18')	14,35	φ0(C18')	14,35	mlogk(C18')	12,82
φ0(C18')	10,61	φ0(C18')	10,61	mlogk(C18')	9,85	mlogk(C18')	15,84	mlogk(C18')	15,84	PC1/logk(C18')	12,82
PC1/logk(C18')	11,36	PC1/logk(C18')	11,36	PC1/logk(C18')	10,61	PC1/logk(C18')	15,84	PC1/logk(C18')	15,84	φ0(C18')	14,32
PC1/logk (C18)	12,88	PC1/logk (C18)	12,88	PC1/logk (C18)	11,36	logkwbin(Ph)	15,85	logkwbin(Ph)	15,85	PC1/logk (C18)	15,09
logkwbin(Ph)	13,26	logkwbin(Ph)	13,26	logkwbin(Ph)	12,50	PC1/logk (C18)	15,88	PC1/logk (C18)	15,88	logkwbin(Ph)	15,84
logkwbin(C18')	13,64	logkwbin(C18')	13,64	logkwbin(C18')	12,88	logkwbin(C18')	16,65	logkwbin(C18')	16,65	φ0(PhF5)	16,63
φ0(Ph)	13,64	φ0(Ph)	13,64	mlogk(Ph)	12,88	φ0(PhF5)	18,14	φ0(PhF5)	18,14	mlogk(Ph)	16,63
mlogk(Ph)	14,39	mlogk(Ph)	14,39	φ0(Ph)	12,88	mlogk(Ph)	18,14	mlogk(Ph)	18,14	PC1/logk(Ph)	16,63
logkwlin(C18')	14,77	logkwlin(C18')	14,77	logkwlin(C18')	13,26	PC1/logk(Ph)	18,14	PC1/logk(Ph)	18,14	logkwbin(C18')	16,60
PC1/logk(Ph)	14,77	PC1/logk(Ph)	14,77	PC1/logk(Ph)	13,26	φ0(Ph)	18,16	φ0(Ph)	18,16	φ0(Ph)	17,38
φ0(PhF5)	15,53	φ0(PhF5)	15,53	φ0(PhF5)	14,02	PC1/logk(PhF5)	18,18	PC1/logk(PhF5)	18,18	PC1/logk(PhF5)	17,40
milogP	16,29	milogP	16,29	PC1/logk(PhF5)	14,77	logkwlin(C18')	18,16	logkwlin(C18')	18,16	logkwlin(C18')	18,14
PC1/logk(PhF5)	16,29	PC1/logk(PhF5)	16,29	mlogk(PhF5)	15,15	mlogk(PhF5)	18,94	mlogk(PhF5)	18,94	mlogk(PhF5)	18,15
logkwbin(PhF5)	16,67	logkwbin(PhF5)	16,67	milogP	16,29	logkwbin(PhF5)	18,93	logkwbin(PhF5)	18,93	logkwbin(PhF5)	21,90
mlogk(PhF5)	16,67	mlogk(PhF5)	16,67	logkwbin(PhF5)	16,67	milogP	20,46	milogP	20,46	milogP	24,25
MLOGP	18,94	MLOGP	18,94	logkwlin(Ph)	18,18	MLOGP	24,94	MLOGP	24,94	MLOGP	27,25
logkwlin(Ph)	18,94	logkwlin(Ph)	18,94	MLOGP	20,45	XLOGP2	31,77	XLOGP2	31,77	logkwlin(Ph)	31,01
AClogP	21,59	AClogP	21,59	AClogP	21,21	logkwlin(Ph)	32,53	logkwlin(Ph)	32,53	AClogP	34,76
logkwlin(PhF5)	21,97	logkwlin(PhF5)	21,97	logkwlin(PhF5)	21,21	AClogP	33,24	AClogP	33,24	logkwlin(PhF5)	35,63
ABlogP	22,35	ABlogP	22,35	ABlogP	21,97	logkwlin(PhF5)	38,63	logkwlin(PhF5)	38,63	log(p)B	37,04
XLOGP2	22,73	XLOGP2	22,73	log(p)B	23,11	log(p)B	39,32	log(p)B	39,32	ABlogP	37,89
log(p)B	23,11	log(p)B	23,11	XLOGP2	23,11	ABlogP	40,89	ABlogP	40,89	XLOGP2	37,78
logP	24,62	logP	24,62	logkwlin (C18)	24,62	Hy	42,29	Hy	42,29	Hy	43,14
logkwlin (C18)	25,38	logkwlin (C18)	25,38	logP	25,38	CLOGP	43,08	CLOGP	43,08	CLOGP	43,13
CLOGP	26,52	CLOGP	26,52	mlogk(C18)	26,52	KowWIN	43,11	KowWIN	43,11	KowWIN	43,16
mlogk(C18)	26,52	mlogk(C18)	26,52	Hy	27,27	ALOGP	44,62	ALOGP	44,62	logkwlin (C18)	44,66
Hy	26,89	Hy	26,89	logDa	28,03	logP	44,67	logP	44,67	mlogk(C18)	45,39
ALOGP	26,89	ALOGP	26,89	CLOGP	28,03	logkwlin (C18)	45,42	logkwlin (C18)	45,42	ALOGP	46,14
logDa	27,27	logDa	27,27	XLOGP3	28,41	mlogk(C18)	45,40	mlogk(C18)	45,40	logP	46,18
KowWIN	27,27	KowWIN	27,27	ALOGP	28,79	logDa	46,18	logDa	46,18	S(Ph)	47,67

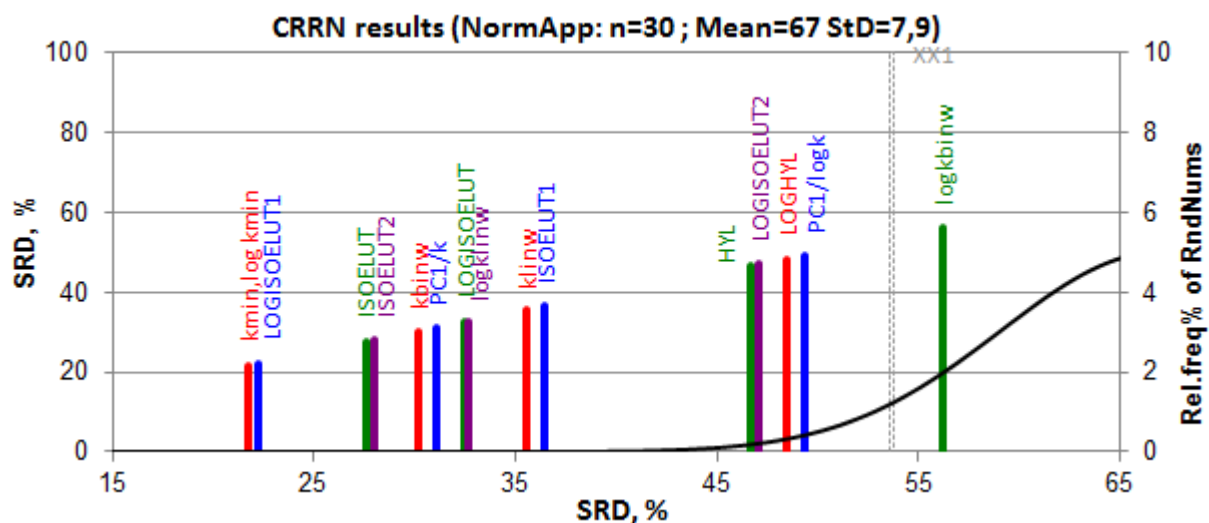
SRD scores						PCM scores (RScale)					
AS			IS			AS			IS		
Variable			Variable			Variable			Variable		
ALOGPs	28,41	ALOGPs	28,41	KowWIN	28,79	log(p)C	47,69	log(p)C	47,69	logDa	48,43
XLOGP3	28,79	XLOGP3	28,79	ALOGPs	29,17	S(Ph)	47,68	S(Ph)	47,68	COSMOFraq	49,90
log(p)C	29,17	log(p)C	29,17	log(p)C	31,06	ALOGPs	48,45	ALOGPs	48,45	S(C18)	50,72
COSMOFraq	33,33	COSMOFraq	33,33	COSMOFraq	32,58	S(C18)	50,70	S(C18)	50,70	log(p)C	50,68
log(p)V	33,71	log(p)V	33,71	S(C18)	34,47	XLOGP3	50,68	XLOGP3	50,68	ALOGPs	50,69
S(C18)	34,47	S(C18)	34,47	S(Ph)	34,47	COSMOFraq	50,67	COSMOFraq	50,67	XLOGP3	51,43
S(Ph)	34,47	S(Ph)	34,47	log(p)V	35,61	log(p)V	53,64	log(p)V	53,64	log(p)V	53,69
logPC	35,61	logPC	35,61	logPC	36,36	S (PhF5)	56,76	S (PhF5)	56,76	logPC	55,26
S (PhF5)	37,12	S (PhF5)	37,12	S (PhF5)	37,12	logPC	56,73	logPC	56,73	S (PhF5)	56,71
S(C18')	52,27	S(C18')	52,27	S(C18')	51,52	S(C18')	64,38	S(C18')	64,38	S(C18')	63,63
logkwbin(C18)	65,91	logkwbin(C18)	65,91	logkwbin(C18)	65,91	logkwbin(C18)	65,91	logkwbin(C18)	65,91	logkwbin(C18)	65,91

One of the referees asked us to separate the different lipophilicity parameters as follows HILIC, HPLC, and Caculated logpPs). Therefore, we carried out a ranking on HILIC lipophilicity parameters using the data of S1b.

Table S7 Case study 1 – SRD ranking of autoscaled HILIC lipophilicity parameters with ties.

Name	kmin	log kmin	LOGISO- ELUT1	ISOELUT	ISO- ELUT2	kbinw	PC1/k	LOGISO- ELUT	logklinw	klinw	ISO- ELUT1	HYL	LOGISO- ELUT2	LOGHYL	PC1/log k	lokkwbin
SRD	98	98	100	124	126	136	140	146	147	160	164	210	212	218	222	253
SRDscaled (0-100)	21.8	21.8	22.2	27.6	28.0	30.2	31.1	32.4	32.7	35.6	36.4	46.7	47.1	48.4	49.3	56.2
probability of	6.31E- 07	6.31E- 07	8.74E-07	3.45E-05	4.59E- 05	1.84E- 04	3.14E- 04	6.83E-04	7.69E-04	3.78E- 03	6.00E- 03	5.23E-01	6.14E-01	0.973	1.31	8.71
random ranking, %	7.33E- 07	7.33E- 07	1.01E-06	3.94E-05	5.24E- 05	2.09E- 04	3.55E- 04	7.69E-04	8.80E-04	4.22E- 03	6.68E- 03	5.65E-01	6.62E-01	1.05	1.40	9.17

The pattern may better be seen in Supplermantary Figure 1:



The ranking should be compared with the original full (complete) SRD ranking (Figure 3). If we eliminate the non-HILIC lipophilicity parameters from figure 3, we receive basically the same pattern: kmin, logkmin, and logisoelut are located ahead (they are the best HILIC lipophilicity parameters), whereas logkbinw, HYL, LogISOELUT2, LogHYL, are ranked as last ones (worst describing HILIC lipophilicity) for pyridinium oxime derivatives. Keeping in mind that the reference (benchmark) is different, the number of parameters averaged is different, the similarities in patterns are better than expected.

References

- [1] V. Voicu, C. Sarbu, F. Tache, F. Micale, S. F. Radulescu, K. Sakurada, H. Ohta, A. Medvedovici, Lipophilicity indices derived from the liquid chromatographic behavior observed under bimodal retention conditions (reversed phase/hydrophilic interaction, Application to a representative set of pyridinium oximes, *Talanta*, 122 (2014) 172-179.
- [2] F. Tachea, R. D. Nascu-Briciu, C. Sarbu, F. Micale, A. Medvedovici, Estimation of the lipophilic character of flavonoids from the retention behavior in reversed phase liquid chromatography on different stationary phases: A comparative study, *J. Pharm. Biomed. Anal.* 57 (2012) 82-93.