Comparison of in-silico modelling and reversed-phase liquid chromatographic retention on an octadecyl silica column to predict skin permeability of pharmaceutical and cosmetic compounds

Grooten, Yasmine; Sych, Aleksandra; Mangelings, Debby; Vander Heyden, Yvan

Published in:
Journal of Pharmaceutical and Biomedical Analysis

DOI:
10.1016/j.jpba.2021.114095

Publication date:
2021

Comparison of in-silico modelling and reversed-phase liquid chromatographic retention on an octadecyl silica column to predict skin permeability of pharmaceutical and cosmetic compounds

Yasmine Grooten, Aleksandra Sych, Debby Mangelings, Yvan Vander Heyden*

Vrije Universiteit Brussel (VUB), Department of Analytical Chemistry, Applied Chemometrics and Molecular Modelling, Laarbeeklaan 103, B-1090 Brussels, Belgium

*: Corresponding author. Tel: +3224774734. E-mail address: yvan.vander.heyden@vub.be

1. Introduction

The measurement of skin permeability of chemicals, i.e. determining the rate at which a compound will penetrate the skin, is of great importance in areas such as drug development and safety assessment, because it indicates whether the compound has the potency to be efficacious or harmful, respectively. Skin permeability can be measured *in vivo* and *in vitro*. However, the introduction of the European ban on animal testing for cosmetics, in accordance with the European Cosmetic Regulation [1], has further increased the necessity for alternative methods. The use of computational models, the so-called Quantitative Structure-Activity Relationship (QSAR) or more specifically Quantitative Structure-Permeability Relationship (QSPR) models [2] may offer a solution. They link the skin permeability to certain structural properties of molecules, so-called molecular descriptors. These models provide an economic alternative with less ethical issues and often provide insight in the mechanisms that permit skin permeation [2,3]. In these models, the skin permeability coefficient K_p (expressed in cm/h), representing the linear velocity of a compound across the skin, is modelled. K_p can also be regarded as the flux at steady-state (J_{ss}) divided by the applied concentration C_v [4]:

$$K_p = J_{ss}/C_v \qquad \text{(Eq. 1)}$$

Flynn [5] was the first to provide a database of log K_p values, measured *in vitro* through human skin and coming from aqueous vehicles, by combining the data from different sources. Potts and Guy [6] were one of the first to model the skin permeability data from this database as a function of the molecular weight (*MW*) and the octanol/water partition coefficient (log *P*):

$$\log K_p = \mathrm{b}_0 + \mathrm{b}_1\, MW + \mathrm{b}_2 \log P \qquad \text{(Eq. 2)}$$

The selected descriptors may thus be related to permeation, since small molecules (low molecular weight) with a more lipophilic character (higher log *P* value) can pass easier through the skin. Since then, many QSPR models have been published, in which other descriptors, such as number of hydrogen bond donors/acceptors, melting points and molecular volumes, were included [7–9]. Several modelling techniques were also applied: linear regression techniques, e.g. multiple linear regression (MLR), and partial least squares (PLS) regression, as well as non-linear regression techniques, such as artificial neural networks (ANN), and classification and regression trees (CART) [10]. In linear free-energy

relationship (LFER) models, a property is modelled by MLR as a function of some specific descriptors [11].

Besides theoretical molecular descriptors, chromatographic retention parameters, which are experimental descriptors, can also be used to model the skin permeability. They result into Quantitative Retention-Activity Relationship (QRAR) models. The retention in different liquid chromatographic methods has already been used to model the skin permeability. Besides regular reversed-phase conditions on C18 columns, more biomimicking conditions have been applied to achieve a better correspondence with the mechanisms of skin permeation. One example is micellar liquid chromatography (MLC), in which a surfactant (e.g. sodium dodecyl sulphate (SDS) or polyoxyethylene-23-lauryl ether (Brij-35)) is added to the mobile phase in a concentration exceeding the critical micellar concentration. Micelles, which resemble the cellular lipid matrix, are then formed in the mobile phase [12–14]. Furthermore, the retention on columns containing certain compounds of the skin, such as keratin [15], collagen [16], cholesterol [17] and phospholipid analogues (immobilized artificial membrane (IAM) columns) [18,19], was previously also applied to model skin permeability. Lázaro et al. [20], for instance, showed a higher correlation of retention on C18 columns compared to IAM columns for modelling skin permeability. These methods were later compared to two micellar electrokinetic chromatographic (MEKC) methods by Hidalgo-Rodríguez et al. [21], and the retention on the C18 columns (with addition of McGowan's volume as descriptor in the model) again provided the best correlation with skin permeability.

The goal of this paper is to explore the added value of an experimental descriptor (log k) in QSPR models. Therefore, the retention of a diverse test set of 58 compounds was measured on a C18 column at two pH levels: pH 7, as in the methods described in the literature [20,21], and additionally at pH 5.5, the pH of the skin. Different fractions of organic modifier were applied to measure retention. They allow the estimation of an extrapolated log k_w value, representing the retention in a purely aqueous mobile phase (without modifier). Consequently, log k_w provides for a given chromatographic system directly comparable values for a diverse set of compounds with an extended range of log P values. Further, two sets of theoretical molecular descriptors were calculated, i.e. with Vega ZZ and E-Dragon software programs, which were also used to model the skin permeability. These theoretical descriptors contain

information on a large series of physicochemical, topological and geometrical properties. Their calculation does not require synthesis of the analytes nor any experimental work. QSPR models were built with and without log *k*, using MLR and PLS as modelling techniques, and their performances were compared to evaluate the potentially added value of the chromatographic descriptor. The use of a chromatographic descriptor could be combined with an *in-silico* approach to model the skin permeability, extending the models containing only the results from a chromatographic approach. In this way, it was also explored whether the contribution of the chromatographic descriptor to the models containing solely theoretical descriptors justifies the experimental work of determining the descriptor. The approach studied will also form a reference to evaluate the added value of other chromatographically determined descriptors, i.e. with other techniques (e.g. SFC) or on other stationary phases.

2. Materials and methods

2.1. Chemicals

Acetonitrile (ACN) and methanol, both HPLC grade, were purchased from VWR Chemicals (Fontenay-sous-Bois, France). Sodium dihydrogenphosphate monohydrate and sodium acetate (both from Sigma Aldrich, Steinheim, Germany) were used for the preparation of the buffers, of which the pH was adjusted with 1 M hydrochloric acid or 1 M sodium hydroxide (both from Fisher Scientific, Loughborough, UK). Ultrapure water was obtained from an Arium Pro UV system (Sartorius Stedim Biotech, Göttingen, Germany).

2.2. Chromatographic conditions

The chromatographic experiments were performed on a Merck-Hitachi HPLC system (Tokyo, Japan) with a quaternary L-7100 pump, an L-7200 autosampler with a 100 µL loop, an L-7400 UV detector and a D-7000 interface. An XTerra RP18 column (150 mm x 4.6 mm i.d., 5 µm) from Waters (Milford, MA, USA) was used as stationary phase. A flow of 1 mL min^{-1} was applied and 10 µL of each sample solution was injected. All compounds were detected at a wavelength of 220 nm at ambient temperature.

The chromatographic results were processed with the D-7000 HPLC System Manager software (Merck-Hitachi, 1994–2001, version 4.1).

The mobile phase consisted of buffer-acetonitrile, at different pH levels. At pH 7, a 10 mM phosphate buffer was used, combined with an acetonitrile fraction, ranging from 25% to 40 % v/v (in steps of 5% v/v). At pH 5.5, mimicking the pH of the skin, a 10 mM acetate buffer was applied, also mixed with fractions of 25% to 45% v/v acetonitrile (in steps of 5% v/v). Buffers were vacuum-filtered through 0.20 µm membranes (Sartorius Stedim Biotech) and before use, all mobile phases were degassed in an ultrasonic bath.

Retention factors k were calculated as $k = (t_r - t_0)/t_0$, with t_r the retention time of the corresponding compound and t_0 the dead time, determined with a 0.1 mg mL^{-1} uracil standard (Fluka, Neu-Ulm, Switzerland) in water.

2.3. Test set

The test set used in this study consisted of 58 compounds, covering a relevant log *P* range from -1.13 to 4.45 and a log K_p range from -5.52 to -0.24. It contains components from different pharmacological classes, such as corticosteroids, anti-inflammatory drugs and hormones; but also preservatives, hair dye agents and antiseptics. An overview of the individual log K_p, log *P*, molecular weight (*MW*) and manufacturer can be found in **Table 1**. The compounds were obtained from Sigma-Aldrich (Steinheim, Germany), Merck (Darmstadt, Germany), Bios Coutelier (Brussels, Belgium), Certa (Braine-l'Alleud, Belgium), Diosynth (Oss, The Nederlands) and Fluka (Neu-Ulm, Switzerland), and have a minimal purity of 95%. The corresponding log K_p versus log *P* values are plotted in **Fig. S1** in the Supplementary material. Solutions with a concentration of 0.1 mg ml^{-1} were prepared in methanol.

2.4. Data sources, software and data processing

Most skin permeability data were obtained from a validated database [22] and originate from *in-vitro* tests. Four compounds from other sources were added [23–26]. Vega ZZ version 3.1.2.29 [27] was used to calculate 21 physicochemical, geometrical and topological descriptors (e.g. number of atoms,

molecular weight, gyration radius, virtual log *P*, surface area and volume). The melting point was taken from PubChem [28]. E-Dragon software was used to calculate 1666 additional molecular descriptors [29], of which the (nearly) constant ones were removed. When descriptor pairs showed a correlation coefficient above 0.95, only the one with the highest correlation to the skin permeability was kept. This reduced the number of applied descriptors to 408. For modelling (in Matlab) with the Vega ZZ descriptors, the number was reduced to 12 after deleting highly correlated descriptors. Autoscaling was also assessed as pre-treatment for the descriptors. In autoscaling, each column of the matrix was centered, after which it was divided by the column standard deviation.

The MLR models were built with the 'automatic linear regression' module of Vega ZZ, varying the number of included descriptors between one and seven. Descriptors with an r^2 below 0.10 with log K_p were omitted for modelling. Furthermore, collinear descriptors, for which the Variance Inflation Factor (VIF = $1/(1-r^2)$) value was above 5.0, were never selected together in a model. In the 'linear regression' module, the descriptors to be included in the model, were selected manually when the chromatographic descriptor showed a low correlation with the skin permeability ($r^2 < 0.10$).

Stepwise MLR and PLS models were obtained by applying m-files written in MATLAB® 2014a (The Mathworks, Natick, MA, USA). The stepwise MLR procedure alternates a forward selection and backward elimination to select the best descriptors for the model. In a first step, the descriptor with the highest correlation to the skin permeability coefficient is selected (evaluating its significance to the model with an overall F-test). Afterwards, descriptors are added (and possibly removed), determining their contribution to the model with a partial F-test. This process is ended when the model can no longer be improved with the addition or removal of descriptors (based on a partial F-test). In the PLS approach, the number of latent variables for the best model is selected based on the lowest root mean squared error of cross-validation (RMSECV) value. The RMSECV from a leave-on-out cross-validation and the root mean squared error of calibration (RMSEC) for all models were also determined with MATLAB. They assess the predictive capabilities and fit of the model, respectively. Low values of these parameters indicate a good model. Relative percentages of the RMSEC and RMSECV were calculated on the average log K_p value.

Microsoft Excel (Microsoft Office Professional Plus 2016, Redmond, WA, USA) was used to calculate the determination coefficient r^2 of each model (between log K_p(experimental) and log K_p(predicted)). The latter was used as another way to look at the predictive capacities of the models. GraphPad Prism (GraphPad Software, San Diego, CA, USA, Version 8.4.3) was used to create the plots.

3. **Results and discussion**

3.1. Chromatographic measurements

With a test set containing compounds covering a broad range of lipophilic properties, it is usually impossible to find one isocratic mobile phase capable of measuring all compounds. Therefore, log k_w, indicating the retention factor at a pure aqueous mobile phase, is estimated. This value is determined by extrapolating the retention factors at different mobile phases using the following equation [30]:

$$\log k = \log k_w - s\,\varphi \qquad \text{(Eq. 3)}$$

in which k and k_w are the retention factors with mixed mobile phases (containing organic modifier) and pure aqueous mobile phase, respectively, while φ represents the fraction organic modifier in the mobile phase. The slope s depends on the compound and the applied chromatographic system.

The measured log k values and extrapolated log k_w values at pH 5.5 and pH 7 can be found in **Tables S1-S2** and **Tables S3-S4** (in the Supplementary material), respectively. All compounds eluted from the column with every tested ACN fraction. At pH 5.5, an additional mobile phase with 45% v/v ACN was tested to obtain retention factors below 10. The last eluting compound at both pH levels was progesterone, with retention times of 148 min at pH 5.5 and 107 min at pH 7, for the lowest fraction of organic modifier (25% v/v ACN). For some components (e.g. thiourea) it was not possible to calculate a log k value, due to no retention (k = 0). Therefore, it was also impossible to calculate a log k_w value for thiourea at pH 5.5 and benzoic acid at pH 7. For most compounds, the determination coefficient (r^2) of Eq. 3 was above 0.90. Nevertheless, compounds with short retention times (and therefore low affinity for the column) seemed more difficult to extrapolate to log k_w, leading to lower r^2-values. This observation is more pronounced at pH 7 than at pH 5.5. In the end, 56 log k values were determined for

the different mobile-phase compositions at pH 5.5 (without thiourea and p-phenylenediamine) and pH 7 (without benzoic acid and p-phenylenediamine), except for 25% and 30% v/v ACN at pH 7, for which only 55 compounds were considered (additionally without thiourea).

Although the difference between the two pH levels was not that pronounced, an effect can be noticed on the retention of certain compounds. A first prominent group of compounds, showing a shift in retention as a result of the pH change, are the nonsteroidal anti-inflammatory drugs (NSAIDs). Several compounds from this group, e.g. diclofenac, flurbiprofen, ibuprofen and indomethacin, contain a carboxylic acid group with a pKa value between 4 and 5. This means that at pH 5.5 these compounds will be partially ionized, while at pH 7, the molecule will be almost fully charged. Consequently, shorter retention times were noticed at pH 7. However, for ketoprofen and naproxen, which contain also a carboxylic acid, and for piroxicam this effect on the retention is less pronounced. 2,4,6-trichlorophenol, containing a phenolic group, showed a similar retention at both pH levels. For lidocaine, a compound with basic properties, an opposite shift can be noticed: at pH 7 the amine group is partially ionized, while at pH 5.5 complete ionization is reached. This leads to a shorter retention of the compound at the lowest pH level. For other compounds, such as chloroxylenol and the steroid hormones (17α-hydroxyprogesterone, estriol, estrone, progesterone, testosterone and β-estradiol) longer retention is noticed at pH 5.5, which can be assigned less to the ionization of the compounds.

3.2. Correlation between log *k* and the skin permeability

The skin permeability coefficients, log K_p, were obtained from *in-vitro* measurements. However, it should be noted that these data are taken from different sources. This is not ideal because of possible interlaboratory differences (preferably the measurements should be performed under a standardized procedure in one laboratory). However, since it is challenging to find large sets of skin permeability data, data from different sources are often combined. To assess the obtained chromatographic data, the correlation between the retention factors, log *k*, and the skin permeability coefficients, log K_p, was checked for each mobile-phase composition and for the extrapolated aqueous mobile phase. Rather poor correlations were noticed, displaying a maximal value (r = 0.317) with 45% v/v ACN at pH 5.5. The correlation between log K_p and the chromatographic data at pH 5.5 was overall slightly better than at

pH 7. This seems logical because 5.5 is the pH of the skin and thus reflects the real skin conditions better. The retention factors from the mobile-phase compositions were mutually highly correlated (r = 0.828 to 0.997 at pH 5.5, and r = 0.716 to 0.989 at pH 7). This means that all retention factor sets contain roughly the same information. However, log k values from the mobile phases with a higher percentage of ACN are better correlated with the skin permeability. On the contrary, the extrapolated log k_w at pH 5.5 showed the lowest correlation with the skin permeability at this pH level, while at pH 7 the correlation with log k_w was one of the highest.

In conclusion, the retention factors demonstrated a low correlation with the log K_p values, rendering them individually inadequate to predict the skin permeability of compounds. Because the skin permeability process is influenced by a combination of different characteristics, this points out that the retention on the C18 column only comprehends part of this process. Thus, the addition of other molecular descriptors may improve the modelling of the skin permeability.

3.3. Modelling skin permeability using only theoretical descriptors

Two sets of molecular descriptors were calculated with Vega ZZ and E-Dragon software programs. The first led to a smaller set of descriptors representing physicochemical, geometrical and topological properties, which may be easier to link to certain skin permeability processes. The set of E-Dragon descriptors was far more extensive, containing also rather abstract theoretical descriptors. Two linear regression techniques were applied to model the skin permeability, log K_p, as a function of the molecular descriptors: MLR and PLS (both on raw and pre-processed descriptors).

The stepwise MLR alternates a forward selection with a backwards deletion of descriptors, resulting in one model based on the fit of the model. The descriptor classes and the selected E-Dragon descriptors in the MLR models can be found in **Table 2**, while **Table 3** shows the best stepwise MLR models with the E-Dragon descriptors. The r^2 value of the stepwise MLR model on the raw data is high, indicating a good fit of the model (see **Fig. 1A**). When autoscaling the data, the same 10 E-Dragon descriptors were selected for the model but here the contribution of every descriptor to the model can directly be assessed comparing their respective coefficients. The first four descriptors showed the largest

influence, with *RDF020e* and *SRW09* having an inversely proportional influence and *C-025* and *RDF055p* a proportional one on log K_p (Eq. 5 **Table 3**). The low RMSEC and RMSECV values confirm the suitability of the stepwise MLR models, showing their good fit and predicative abilities.

For the PLS regression with the E-Dragon descriptors, the best model was selected as that with a number of PLS factors that showed the lowest RMSECV value. Without data pre-treatment, the best model contained 6 PLS factors, having an RMSEC value of 0.674 (or 25.1%, calculated on the average log K_p value), an RMSECV of 0.860 (or 32.0%) and an r^2 of 0.72. When autoscaling the descriptors, 7 PLS factors were selected, leading to an overall improvement of the model: an RMSEC value of 0.292 (or 10.9%), RMSECV of 0.740 (or 27.5%) and r^2 of 0.95. Compared to the stepwise MLR models, the PLS model without data pre-treatment showed a less good fit. Considering the much higher RMSEC and RMSECV values, this model was thus less good. Though the PLS model built with the autoscaled descriptors showed an improvement in RMSEC and r^2 values (see also **Fig. 1B)**, the RMSECV value was still fairly high. The latter is visualized in **Fig. 2**, showing the leave-one-out cross-validated log K_p values, as a function of the experimental. The model tends also to overestimate the permeability of compounds with low experimental log K_p values. In conclusion, stepwise MLR resulted in the preferred regression model for the E-Dragon descriptor set.

The stepwise MLR models with the Vega ZZ descriptors can be consulted in **Table 4**. The influence of the descriptors was assessed from their coefficients in the model with the autoscaled data (Eq. 7). Both descriptors showed a somewhat similar influence on the skin permeability, though their effect was opposite. The RMSEC and RMSECV were, however, quite substantial, thus worse than from the stepwise MLR model with the E-Dragon descriptors.

Secondly, using the automatic linear regression modelling in Vega ZZ, models were built with a pre-selected number of descriptors (ranging from one to seven). This allowed selecting manually the overall best model (making a compromise between its fit and predictive abilities). An overview of these models can be found in **Table 5**. When increasing the number of descriptors, an improvement of the RMSEC and initially of the RMSECV values was noticed. However, at a certain complexity, the models became susceptible to overfitting, indicated by the increase of RMSECV. Therefore, a compromise must

be made between the fit of the model (RMSEC) and its predictive properties (RMSECV), keeping the model as simple as possible. With these conditions in mind, the model with four descriptors (Eq. 11) was selected as the optimal. It had an RMSEC value of 25.7% and an RMSECV of 28.0%, calculated on the average log K_p. It thus corresponded with the results from the other modelling approaches on the Vega ZZ descriptors. The influence of the descriptors was evaluated from the model with the four Vega ZZ descriptors of Eq. 11 but now autoscaled (and having the same statistical parameters):

$$\log K_p = -2.69 + 0.52\ \textit{Virtual log P} + 0.21\ \textit{Lipole (Broto)} - 0.26\ \textit{HbDon} - 0.61\ \textit{Atoms} \quad \text{(Eq. 15)}$$

The virtual log *P* and number of atoms had the largest influence on the skin permeability, while the coefficients of Broto's lipole (calculated according to [31]) and the number of hydrogen bond donors were slightly smaller. It was noticed from **Table 5**, that the RMSEC value of the models with three or more descriptors further improved compared to the model with two descriptors (Eq. 9, equal to the model selected with the stepwise MLR approach, Eq. 6). It can therefore be questioned if the stopping criterion of the stepwise MLR approach is perhaps too strict, seeing that the RMSECV value also improved when expending the model with two additional descriptors.

Additionally, PLS regression was applied on the raw Vega ZZ data, obtaining RMSEC values of 0.757 (or 28.2%) and RMSECV of 0.807 (or 30.0%). Five PLS factors provided the best model with the lowest RMSECV and an r^2 of 0.63. Autoscaling the descriptors, gave rise to a PLS model (3 PLS factors) with slightly better parameters: r^2 = 0.70, RMSEC = 0.682 (or 25.4%), and RMSECV = 0.790 (or 29.4%). These parameters were quite similar to those of the previously obtained models using the same descriptors. Both for calibration and cross-validation the error was quite high. The E-Dragon-based PLS models again seemed superior to those derived from the Vega ZZ descriptors.

3.4. Modelling skin permeability with both theoretical and chromatographic descriptors

To optimize the models which include the chromatographic descriptors, theoretical molecular descriptors were added to the skin permeability models. Models were built for all mobile-phase compositions combining the measured log *k* values with the theoretical descriptors from Vega ZZ, on the one hand, and the E-Dragon descriptors, on the other. The extrapolated log k_w values were also

considered theoretical descriptors for modelling. An overview of the models (**Tables S5 – S18**) was included in the Supplementary material, while the best models were discussed below.

Regular stepwise MLR modelling was not successful, since the retention factors were never added to the model (the same models as before were obtained). Thus, the retention factors did not contribute significantly to these MLR models. However, when excluding the virtual log *P* from the data set with the Vega ZZ descriptors, stepwise MLR models with the log *k* values were obtained for some mobile-phase compositions, of which the extrapolated log k_w at pH 5.5 provided the best model (**Table S5**). Substitution of log *P* by chromatographic retention was done because the latter can be used to estimate log *P* [32].

$$\log K_p = -1.61 + 0.42 \log k_{w.\,pH\,5.5} - 0.053\,Atoms - 0.39\,HbDon + 0.25\,Lipole\,(Broto) \quad \text{(Eq. 16)}$$

RMSEC = 0.722, RMSECV = 0.792, r^2 = 0.67, n = 56

Compared to the model containing four theoretical Vega ZZ descriptors (Eq. 11), the virtual log *P* was substituted by the retention factor. The obtained model (Eq. 16) showed slightly less favourable statistical parameters (RMSEC of 27.3% and RMSECV of 29.9%, calculated on the average log K_p) than the model with only theoretical descriptors. The virtual log *P* is therefore the preferred variable instead of the chromatographic descriptor.

As a second approach, the automatic linear regression models were built using the Vega ZZ software, combining the retention factors and Vega ZZ descriptors. However, for this procedure, a variable needs to have an r^2 value above 0.10 with the log K_p to be considered in the construction of the skin permeability models. This was only the case for the retention factors obtained with 45% v/v ACN at pH 5.5 (see **Table S6** in the Supplementary material). For the other mobile-phase compositions, MLR models were built, manually selecting the same descriptors as with the log $k_{0.45,\,pH\,5.5}$ models (**Table S6**) or adding the log *k* to the best obtained model with only theoretical Vega ZZ descriptors (**Table 5**). With the latter models, the aim was to verify whether the addition of the retention factor could improve the models that contained only theoretical descriptors. The best model was always selected based on a compromise between the fit of the model (RMSEC) and its predictive capacities (RMSECV). When

comparing the skin permeability models obtained at the different mobile phases at pH 5.5 (see **Tables S6 – S11**), the following model, containing the extrapolated log k_w, resulted in the best statistical values (see **Fig. 1C** and Eq. 17):

$$\log K_p = -1.86 + 0.26 \log k_{w\,pH\,5.5} - 0.051\ Atoms + 0.37\ Virtual\ log\ P - 0.22\ HbDon \qquad \text{(Eq. 17)}$$

RMSEC = 0.699 (or 26.4%), RMSECV = 0.774 (or 29.3%), $r^2 = 0.69$, $n = 56$

The model seems also to overestimate the permeability of compounds with low log K_p values. It should be noticed that the difference between the models with the different fractions of ACN was very small, especially with the model containing the log $k_{0.35,\,pH\,5.5}$ (same selected descriptors with RMSEC = 0.704 or 26.6%, RMSECV = 0.771 or 29.1% and $r^2 = 0.68$, see **Table S8**).

At pH 7 (**Tables S12 - S16**), the best models were obtained with the retention factors from 35% v/v ACN, log $k_{0.35,\,pH\,7}$, (**Table S13**) and the extrapolated retention factor, log $k_{w,\,pH\,7}$, (**Table S16**) combined with three Vega ZZ descriptors (Eqs. 18 and 19):

$$\log K_p = -1.71 - 0.11 \log k_{0.35,\,pH\,7} + 0.65\ Virtual\ log\ P - 0.58\ Gyration\ radius - 0.0038\ Melting\ point \qquad \text{(Eq. 18)}$$

RMSEC = 0.701 (or 26.1%), RMSECV = 0.777 (or 28.9%), $r^2 = 0.69$, $n = 56$

$$\log K_p = -2.06 + 0.17 \log k_{w,\,pH\,7} - 0.046\ Atoms + 0.45\ Virtual\ log\ P - 0.20\ HbDon \qquad \text{(Eq. 19)}$$

RMSEC = 0.703 (or 26.1%), RMSECV = 0.770 (or 28.6%), $r^2 = 0.68$, $n = 56$

For both pH levels, the best MLR models were thus obtained combining the log k values with three descriptors. At pH 5.5 (Eq. 17) and for the log k_w at pH 7 (Eq. 19), the number of atoms (*Atoms*), virtual log *P* and number of hydrogen bond donors (*HbDon*) provided the best models. For the other mobile-phase compositions at pH 7 (see **Tables S12 – S15**), the best MLR model included the virtual log *P*, gyration radius and melting point (as selected with the theoretical MLR model in Eq. 10). Similar statistical parameters are noticed between these models at both pH levels.

When comparing these MLR models (Eqs. 17 – 19) to the models containing only theoretical Vega ZZ descriptors (**Table 5**), little improvement is seen relative to the RMSEC of the model

containing three descriptors (Eq. 10). Moreover, the RMSECV is less good for the models including log *k*, making these predictions less reliable. The model with four theoretical Vega ZZ descriptors (Eq. 11, containing Broto's lipole instead of log *k* in Eqs. 17 and 19) was superior to these models both in terms of RMSEC and RMSECV. It can thus be concluded that the chromatographic descriptors provided little contribution to the MLR models with theoretical descriptors.

The log *k* values were thereafter also used to acquire PLS models (**Tables S17 – S18**). The chromatographic descriptor was combined with the E-Dragon descriptors (**Table S17**). The best PLS model for each pH level was obtained with the log k_w (6 PLS factors based on the lowest RMSECV value). For pH 5.5, this provided a model with an RMSEC of 0.670 (25.3%), RMSECV of 0.887 (33.5%) and r^2 of 0.72, while for pH 7 the RMSEC was 0.672 (25.0%), RMSECV 0.870 (32.3%) and r^2 0.72. These values resemble closely the statistical parameters of the PLS model with only E-Dragon descriptors, demonstrating that the retention factors have little to no influence on this PLS model. When using the Vega ZZ descriptors, the best compromises between the fit and predictive capacities of the PLS models were observed using also the log $k_{0.40,\ pH\ 5.5}$ (see **Table S18** and **Fig. 1D**) or the log $k_{w,\ pH\ 7}$, both selecting 6 PLS factors. Somewhat similar statistical parameters were obtained for pH 5.5 (RMSEC = 0.699 or 26.4%, RMSECV = 0.811 or 30.7%, and r^2 = 0.69) and for pH 7 (RMSEC = 0.701 or 26.0%, RMSECV = 0.818 or 30.4%, and r^2 = 0.69). In this case, an improvement of the PLS model was noticed in regards to the PLS model containing only Vega ZZ descriptors (selecting 5 PLS factors). The model in **Fig. 1D** still overestimated the permeability of compounds with a low log K_p.

In conclusion, the best MLR and PLS models combining the theoretical descriptors with the log *k* values obtained at pH 5.5 showed large resemblances to those combined with the chromatographic descriptors from pH 7. The addition of the extrapolated log k_w values was useful since this variable was often included in (one of) the best models. However, often the models including the log *k* obtained with an experimentally tested mobile phase gave similar results, while for the extrapolation much more experimental work needs to be performed (at least three mobile-phase compositions were tested). Furthermore, the use of the log *k* values in general can be questioned, since the addition of the chromatographic descriptor rarely led to an improvement to the models with solely theoretical

descriptors. Therefore, there is no justification for the extensive experimental work, when the same quality of models can be obtained by simply using theoretical descriptors. Only when the retention factors were combined with the Vega ZZ descriptors in PLS models, a positive effect on the statistical parameters of the models was noticed. However, this achieved improvement was rather small and therefore perhaps insufficient to justify the experimental work. Furthermore, the MLR model containing four Vega ZZ descriptors (Eq. 11) showed similar statistical parameters to this PLS model.

4. Conclusions

In this paper, the application of theoretical descriptors for the estimation of skin permeability was examined. In this regard, two sets of molecular descriptors (obtained with E-Dragon and Vega ZZ software) were applied to build MLR and PLS models. The interpretation of the E-Dragon descriptors in relation to physicochemical properties of the molecule was not always straightforward, as can be expected for many theoretical descriptors. In contrast, the physicochemical descriptors from Vega ZZ were often easier to link to properties influencing the skin permeability of compounds and in this way, could perhaps give better insights in processes involved in the permeation of the skin. It was therefore no surprise that properties, such as log *P*, number of hydrogen bond donors and number of atoms, which are known to influence the skin permeability, were selected in the MLR models. The overall best model in this study was the stepwise MLR model including ten E-Dragon descriptors. However, for predictive purposes the model should be further validated with the use of an external test set, though this is quite challenging since there is not a lot of skin permeability data available. Overall, the other MLR and PLS models were of a similar quality relative to each other, showing less good fit and predictions. Since the error on the *in-vitro* measurements of skin permeability can be quite substantial (e.g. 0.43 logarithmic units [24]), the RMSEC and RMSECV of these models seem acceptable for most applications.

The relevance of a chromatographic descriptor to predict skin permeability was evaluated using the retention factors obtained with different fractions of organic modifier at pH 5.5 and pH 7 on a C18 column. The skin permeability was better correlated with the results from pH 5.5. This seems reasonable, considering that 5.5 is the pH of the skin. However, the log *k* values alone were insufficient to predict

the skin permeation, requiring the addition of other descriptors to further improve the model. Often the extrapolated log k_w is included in such models, combining the chromatographic descriptor with other descriptors. However, this approach is far more elaborate (demanding more measuring points) than using the retention factors from one single mobile-phase composition (when possible to measure all compounds, which often is not the case). Furthermore, the retention at different fractions frequently resulted in similar models. Even the distinction between the models at the two pH levels was not very pronounced. However, only for the PLS models using the Vega ZZ descriptors an improvement was seen when the chromatographic descriptor was added to the data set.

In conclusion, the chromatographic descriptors obtained on the C18 column do not show important improvements to most models. In the majority of the approaches, similar or even slightly better models were obtained using only theoretical descriptors. The interactions that affect the retention on the C18 column, do not seem to correspond extensively with the mechanisms involved in skin permeation. The small improvement seen in some PLS models does not justify the extent of the experimental work, when *in-silico* modelling can provide equally or more useful models.

Acknowledgments

This work was supported by the Research Foundation – Flanders (FWO) [grant 11D3518N].

Conflict of Interest

The authors declare no conflict of interest.

Tables

Table 1. The 58 compounds in the test set with their skin permeability coefficient (log K_p), molecular weight (*MW*), octanol-water partition coefficient (log *P*) and manufacturer.

Compound	Log K_p [a]	*MW* [b]	Log *P* [c]	Manufacturer
17α-hydroxyprogesterone	-3.22	330.46	1.84	Sigma-Aldrich
2,4,6-Trichlorophenol	-1.23	197.45	3.94	Sigma-Aldrich
2,4-Dichlorophenol	-1.22	163.00	3.18	Sigma-Aldrich
2-Amino-4-nitrophenol	-3.18	154.12	1.39	Sigma-Aldrich
2-Nitro-p-phenylenediamine	-3.30	153.14	0.78	Sigma-Aldrich
4-Amino-2-nitrophenol	-2.55	154.12	1.22	Sigma-Aldrich
Acetylsalicylic acid	-2.14 [24]	180.16	1.10	Fluka
Aminopyrine	-2.99	231.29	1.57	Sigma-Aldrich
Amylobarbital	-2.64	226.27	1.97	Bios
Antipyrine	-4.18	188.23	1.26	Unknown origin
Atropine	-5.07	289.37	2.14	Sigma-Aldrich
Barbital	-3.96	184.19	0.95	Bios Coutelier
Benzoic acid	-1.52	122.12	1.24	Merck
Benzyl alcohol	-2.22	108.14	1.34	Sigma-Aldrich
Caffeine	-2.80	194.19	-0.25	Fluka
Chloroxylenol	-1.23	156.61	3.06	Sigma-Aldrich
Chlorpheniramine (maleate)	-2.66	274.79	3.86	Sigma-Aldrich
Cortexolone	-4.12	346.46	0.97	Sigma-Aldrich
Cortexone	-3.35	330.46	1.92	Sigma-Aldrich
Corticosterone	-3.19	346.46	0.82	Sigma-Aldrich
Cortisone	-5.00	360.44	-0.12	Sigma-Aldrich
Diclofenac	-1.74	296.15	4.45	Sigma-Aldrich
Ephedrine.HCl	-2.22	165.23	1.73	Sigma-Aldrich
Estriol	-4.40	288.38	2.30	Unknown origin
Estrone	-2.44	270.37	3.18	Diosynth
Ethyl nicotinate	-2.20	151.16	1.35	Sigma-Aldrich
Flurbiprofen	-0.34	244.26	3.88	Sigma-Aldrich
Haloperidol	-4.04 [25]	375.86	3.75	Unknown origin
Hydrocortisone	-5.52	362.46	0.01	Certa
Ibuprofen	-0.24	206.28	3.20	Sigma-Aldrich
Indomethacin	-1.30	357.79	4.04	Sigma-Aldrich
Ketoprofen	-1.23	254.28	2.23	Sigma-Aldrich
Lidocaine	-1.70 [26]	234.34	3.01	Sigma-Aldrich
m-Cresol	-1.82	108.14	2.09	Sigma-Aldrich
Methyl nicotinate	-2.49	137.14	0.93	Sigma-Aldrich
Methyl-4-hydroxybenzoate	-2.04	152.15	1.36	Fluka
m-Nitrophenol	-2.25	139.11	2.21	Sigma-Aldrich
Naproxen	-1.42	230.26	3.23	Sigma-Aldrich
o-Chlorophenol	-1.48	128.56	2.40	Sigma-Aldrich
o-Cresol	-1.80	108.14	2.07	Sigma-Aldrich
Paracetamol	-3.35	151.16	1.07	Sigma-Aldrich
p-Cresol	-0.92	108.14	2.16	Sigma-Aldrich
Phenobarbitone	-3.35	232.24	1.49	Unknown origin
Phenol	-1.71	94.11	1.59	Merck
Piroxicam	-2.47	331.35	1.59	Sigma-Aldrich
p-Nitrophenol	-2.25	139.11	2.17	Sigma-Aldrich
p-Phenylenediamine	-3.62	108.14	0.23	Sigma-Aldrich
Prednisolone	-4.35	360.44	-0.45	Sigma-Aldrich

Progesterone	-2.82	314.46	2.85	Sigma-Aldrich
Resorcinol	-3.62	110.11	1.18	Merck
Salicylic acid	-2.20	138.12	0.96	Sigma-Aldrich
Testosterone	-3.40	288.42	2.48	Sigma-Aldrich
Thiourea	-4.02 [23]	76.12	-0.65	Merck
Thymol	-1.28	150.22	3.23	Sigma-Aldrich
Triamcinolone	-5.40	394.43	-1.13	Sigma-Aldrich
Triamcinolone acetonide	-4.69	434.50	1.40	Sigma-Aldrich
β-estradiol	-2.37	272.38	3.37	Unknown origin
β-naphthol	-1.55	144.17	2.61	Merck

[a] Logarithm of the skin permeability coefficient (log K_p in cm h^{-1}), taken from [22], unless indicated differently,

[b] MW (g mol^{-1}), and [c] virtual log P, calculated with Vega ZZ, version 3.1.1.42 software [27].

Table 2. The selected E-Dragon descriptors in the stepwise MLR models.

Descriptor	**Descriptor class**	**Descriptor name** [33,34]
BEHm1	Burden eigenvalues	Largest eigenvalue n. 1 of Burden matrix weighted by mass
C-025	Atom-centred fragments	R-CR-R
D/Dr06	Ring descriptors	Distance/detour ring index of order 6
G1u	WHIM descriptors	1st component symmetry directional WHIM index / unweighted
JGI4	2D autocorrelations	Mean topological charge index of order 4
Mor26u	3D-MoRSE descriptors	Signal 26 / unweighted
Mor32e	3D-MoRSE descriptors	Signal 32 / weighted by Sanderson electronegativity
RDF020e	RDF descriptors	Radial Distribution Function - 020 / weighted by Sanderson electronegativity
RDF055p	RDF descriptors	Radial Distribution Function - 055 / weighted by polarizability
SRW09	Walk and path counts	Self-returning walk count of order 9

Table 3. The stepwise MLR models built with the E-Dragon descriptors, together with their statistical parameters: root mean squared error of calibration, RMSEC, and of cross-validation, RMSECV, (along with their relative value, calculated on the average log K_p value) and determination coefficient r^2. The descriptors are defined in **Table 2**.

Pre-treatment	**RMSEC**	**RMSECV**	r^2	**Equation**
No pre-treatment	0.378	0.452	0.91	$\log K_p$ = -7.64 - 0.34 *RDF020e* + 0.77 *C-025* + 0.26 *RDF055p* - 0.0012 *SRW09* - 1.20 *G1u* - 0.0086
	(14.1%)	(16.8%)		*D/Dr06* + 5.95 *JGI4* + 0.73 *Mor26u* + 1.63 *BEHm1* + 0.72 *Mor32e* (Eq. 4)
Autoscaling	0.378	0.452	0.91	$\log K_p$ = -2.67 - 0.98 *RDF020e* + 0.73 *C-025* + 1.05 *RDF055p* - 0.85 *SRW09* - 0.25 *G1u* - 0.47
	(14.1%)	(16.8%)		*D/Dr06* + 0.16 *JGI4* + 0.19 *Mor26u* + 0.21 *BEHm1* + 0.16 *Mor32e* (Eq. 5)

Table 4. The stepwise MLR models using the Vega ZZ descriptors, together with their statistical parameters root mean squared error of calibration, RMSEC, and of cross-validation, RMSECV, (along with their relative value, calculated on the average log K_p value) and determination coefficient r^2.

Pre-treatment	RMSEC	RMSECV	r^2	Equation	
No pre-treatment	0.731 (27.7%)	0.768 (28.6%)	0.66	log K_p = -1.66 + 0.73 *Virtual log P* - 0.82 *Gyration radius*	(Eq. 6)
Autoscaling	0.731 (27.7%)	0.768 (28.6%)	0.66	log K_p = -2.69 + 0.92 *Virtual log P* - 0.67 *Gyration radius*	(Eq. 7)

Table 5. The performance parameters of the automatic regression MLR models with the Vega ZZ descriptors, and with increasing complexity.

Nr	RMSEC	RMSECV	r^2	Equation	
1	0.971	1.000	0.39	log K_p = -1.26 - 0.011 *Melting point*	(Eq. 8)
2	0.731	0.768	0.66	log K_p = -1.66 + 0.73 *Virtual log P* - 0.82 *Gyration radius*	(Eq. 9)
3	0.706	0.758	0.68	log K_p = -1.61 + 0.64 *Virtual log P* - 0.61 *Gyration radius* - 0.0036 *Melting point*	(Eq. 10)
4	0.690	0.752	0.69	log K_p = -2.26 + 0.42 *Virtual log P* + 0.19 *Lipole (Broto)* - 0.25 *HbDon* - 0.040 *Atoms*	(Eq. 11)
5	0.677	0.752	0.70	log K_p = -2.51 + 0.50 *Virtual log P* + 0.24 *Lipole (Broto)* - 0.40 *HbDon* + 0.012 *PSA* - 0.0080 *Volume*	(Eq. 12)
6	0.657	0.754	0.72	log K_p = -2.59 + 0.48 *Virtual log P* + 0.26 *Lipole (Broto)* - 0.38 *HbDon* + 0.017 *PSA* - 0.0065 *Volume* - 0.0038 *Melting point*	(Eq. 13)
7	0.656	0.770	0.72	log K_p = -2.63 + 0.47 *Virtual log P* + 0.28 *Lipole (Broto)* - 0.45 *HbDon* + 0.022 *PSA* - 0.0059 *Volume* - 0.0041 *Melting point* - 0.088 *HbAcc*	(Eq. 14)

HbDon = number of hydrogen bond donors, *PSA* = polar surface area, *HbAcc* = number of hydrogen bond acceptors

Figures

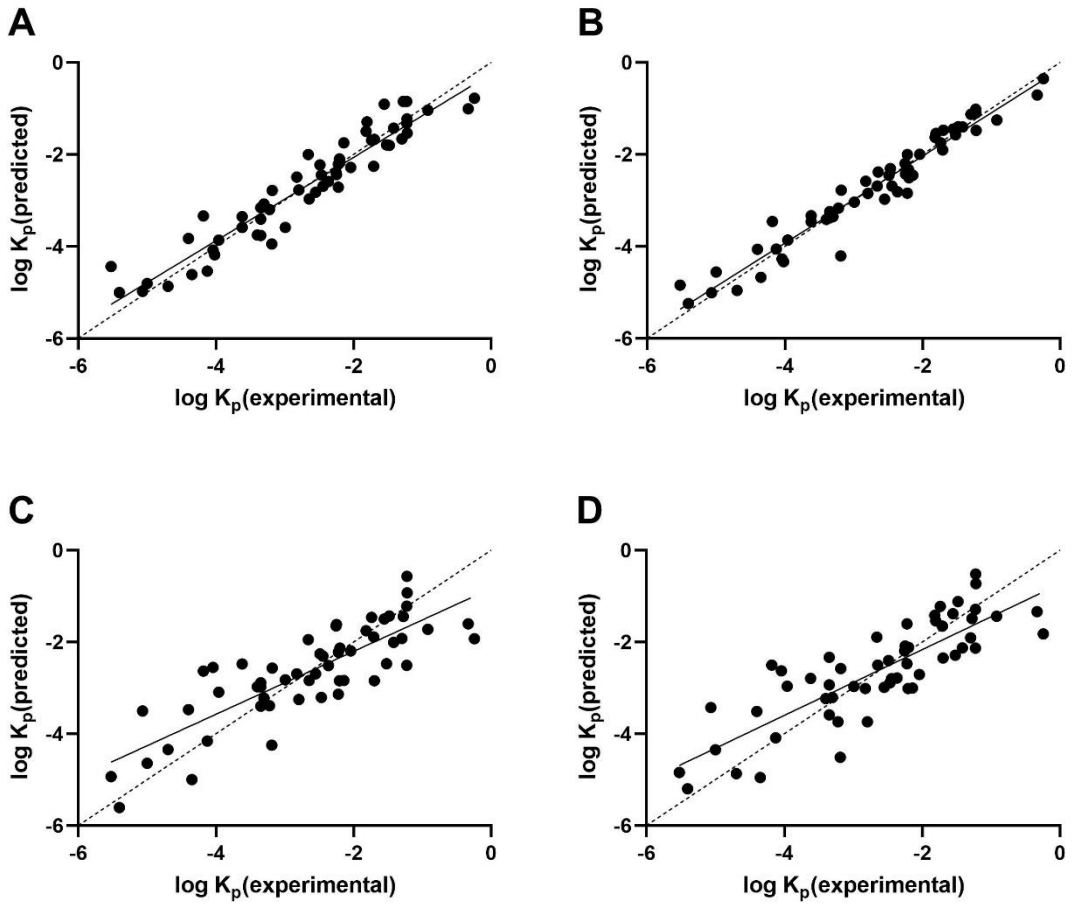

Fig. 1. The predicted log K_p as a function of the experimental, together with the regression line (solid line) and the bisector (dashed line, y = x) for A. the stepwise MLR model containing only E-Dragon descriptors without pre-treatment (Eq. 4); B. the PLS model containing the autoscaled E-Dragon descriptors; C. the MLR model containing the log $k_{w,\,pH\,5.5}$ and three Vega ZZ descriptors (Eq. 17); and D. the PLS model containing the log $k_{0.40,\,pH\,5.5}$ and the Vega ZZ descriptors.

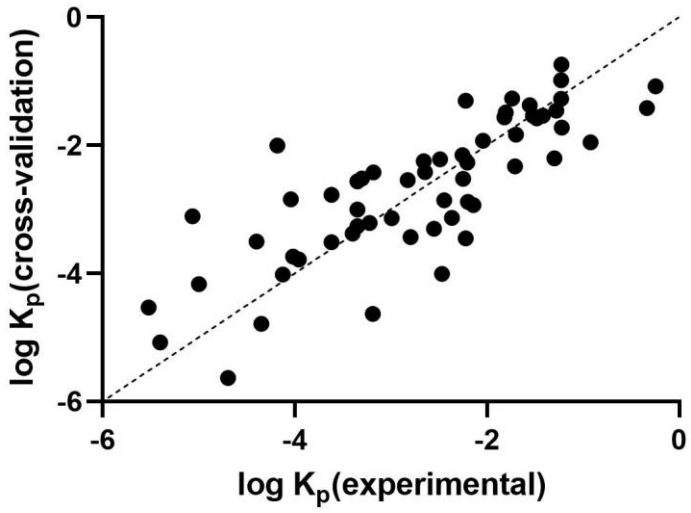

Fig. 2. The log K_p values predicted from leave-one-out cross-validation versus the experimental values using the PLS model with the autoscaled E-Dragon descriptors. The bisector is also drawn.

References

[1] J. Buzek, B. Ask, Regulation (EC) No 1223/2009 of the European Parliament and of the Council of 30 November 2009 on cosmetic products, Off. J. Eur. Union L. 342 (2009) 59–209.

[2] G.P. Moss, J.C. Dearden, H. Patel, M.T.D. Cronin, Quantitative structure – permeability relationships (QSPRs) for percutaneous absorption, Toxicol. Vitr. 16 (2002) 299–317. doi:https://doi.org/10.1016/S0887-2333(02)00003-6.

[3] S. Geinoz, R.H. Guy, B. Testa, P.A. Carrupt, Quantitative Structure-Permeation Relationships (QSPeRs) to Predict Skin Permeation: A Critical Evaluation, Pharm. Res. 21 (2004) 83–92. doi:https://doi.org/10.1023/B:PHAM.0000012155.27488.2b.

[4] J. Kielhorn, S. Melching-Kollmuß, I. Mangelsdorf, Environmental Health Criteria 235: Dermal Absorption, World Health Organization Press, Geneva, 2006.

[5] G.L. Flynn, Physicochemical Determinants of Skin Absorption, in: T.R. Gerrity, C.J. Henry (Eds.), Principles of Route-to-Route Extrapolation of Risk Assessment, Elsevier, New York, USA, 1990: pp. 93–127.

[6] R.O. Potts, R.H. Guy, Predicting Skin Permeability, Pharm. Res. 09 (1992) 663–669. doi:https://doi.org/10.1023/A:1015810312465.

[7] M.D. Barratt, Quantitative structure-activity relationships for skin permeability, Toxicol. Vitr. 9 (1995) 27–37. doi:https://doi.org/10.1016/0887-2333(94)00190-6.

[8] E.J. Lien, H. Gaot, QSAR Analysis of Skin Permeability of Various Drugs in Man as Compared to in Vivo and in Vitro Studies in Rodents, Pharm. Res. An Off. J. Am. Assoc. Pharm. Sci. 12 (1995) 583–587. doi:https://doi.org/10.1023/A:1016266316100.

[9] L.A. Kirchner, R.P. Moody, E. Doyle, R. Bose, J. Jeffery, I. Chu, The Prediction of Skin Permeability by Using Physicochemical Data, Altern. to Lab. Anim. 25 (1997) 359–370. doi:https://doi.org/10.1177/026119299702500319.

[10] I. Tsakovska, I. Pajeva, M. Al Sharif, P. Alov, E. Fioravanzo, S. Kovarich, A.P. Worth, A.N. Richarz, C. Yang, A. Mostrag-Szlichtyng, M.T.D. Cronin, Quantitative structure-skin permeability relationships, Toxicology. 387 (2017) 27–42. doi:https://doi.org/10.1016/j.tox.2017.06.008.

[11] M.H. Abraham, F. Martins, Human skin permeation and partition: General linear free-energy relationship analyses, J. Pharm. Sci. 93 (2004) 1508–1523. doi:https://doi.org/10.1002/jps.20070.

[12] J.J. Martínez-Pla, Y. Martín-Biosca, S. Sagrado, R.M. Villanueva-Camañas, M.J. Medina-Hernández, Biopartitioning micellar chromatography to predict skin permeability, Biomed. Chromatogr. 17 (2003) 530–537. doi:https://doi.org/10.1002/bmc.281.

[13] L.J. Waters, Y. Shahzad, J. Stephenson, Modelling skin permeability with micellar liquid chromatography, Eur. J. Pharm. Sci. 50 (2013) 335–340. doi:https://doi.org/10.1016/j.ejps.2013.08.002.

[14] V. Dobričić, K. Nikolic, S. Vladimirov, O. Čudina, Biopartitioning micellar chromatography as a predictive tool for skin and corneal permeability of newly synthesized 17β-carboxamide steroids, Eur. J. Pharm. Sci. 56 (2014) 105–112. doi:https://doi.org/10.1016/j.ejps.2014.02.007.

[15] M. Turowski, R. Kaliszan, Keratin immobilized on silica as a new stationary phase for chromatographic modelling of skin permeation, J. Pharm. Biomed. Anal. 15 (1997) 1325–1333. doi:https://doi.org/10.1016/S0731-7085(96)02009-2.

[16] M. Turowski, R. Kaliszan, Collagen immobilised on silica derivatives as a new stationary phase for HPLC, Biomed. Chromatogr. 12 (1998) 187–192. doi:https://doi.org/10.1002/(SICI)1099-0801(199807/08)12:4<187::AID-BMC727>3.0.CO;2-2.

[17] M. Janicka, M. Sztanke, K. Sztanke, Reversed-phase liquid chromatography with octadecylsilyl, immobilized artificial membrane and cholesterol columns in correlation studies

with in silico biological descriptors of newly synthesized antiproliferative and analgesic active compounds, J. Chromatogr. A. 1318 (2013) 92–101. doi:https://doi.org/10.1016/j.chroma.2013.09.060.

[18] F. Barbato, B. Cappello, A. Miro, M.I. La Rotonda, F. Quaglia, Chromatographic indexes on immobilized artificial membranes for the prediction of transdermal transport of drugs, Farm. 53 (1998) 655–661. doi:https://doi.org/10.1016/S0014-827X(98)00082-2.

[19] A. Nasal, M. Sznitowska, A. Buciński, R. Kaliszan, Hydrophobicity parameter from high-performance liquid chromatography on an immobilized artificial membrane column and its relationship to bioactivity, J. Chromatogr. A. 692 (1995) 83–89. doi:https://doi.org/10.1016/0021-9673(94)00689-7.

[20] E. Lázaro, C. Ràfols, M.H. Abraham, M. Rosés, Chromatographic estimation of drug disposition properties by means of immobilized artificial membranes (IAM) and C18 columns, J. Med. Chem. 49 (2006) 4861–4870. doi:https://doi.org/10.1021/jm0602108.

[21] M. Hidalgo-Rodríguez, S. Soriano-Meseguer, E. Fuguet, C. Ràfols, M. Rosés, Evaluation of the suitability of chromatographic systems to predict human skin permeation of neutral compounds, Eur. J. Pharm. Sci. 50 (2013) 557–568. doi:https://doi.org/10.1016/j.ejps.2013.04.005.

[22] B.E. Vecchia, A.L. Bunge, Skin absorption databases and predictive equations, in: R.H. Guy, J. Hadgraft (Eds.), Transdermal Drug Delivery, Marcel Dekker, New York, USA, 2003: pp. 57–141.

[23] K.T. Hoang, Dermal Exposure Assessment: Principles and Applications, Environmental Protection Agency, Office of Health and Environmental Assessment, EPA/600/8–91/011B, Washington, DC, USA, 1992.

[24] I.T. Degim, W.J. Pugh, J. Hadgraft, Skin permeability data: anomalous results, Int. J. Pharm. 170 (1998) 129–133. doi:https://doi.org/10.1016/S0378-5173(98)00113-6.

[25] A.F. Azarbayjani, H. Lin, C.W. Yap, Y.W. Chan, S.Y. Chan, Surface tension and wettability in transdermal delivery: a study on the in-vitro permeation of haloperidol with cyclodextrin across human epidermis, J Pharm Pharmacol. 62 (2010) 770–778. doi:https://doi.org/10.1211/jpp.62.06.0014.

[26] J. Thomas, S. Majumdar, S. Wasdo, A. Majumdar, K.B. Sloan, The effect of water solubility of solutes on their flux through human skin in vitro: an extended Flynn database fitted to the Roberts-Sloan equation, Int J Pharm. 339 (2007) 157–167. doi:https://doi.org/10.1016/j.ijpharm.2007.02.031.

[27] A. Pedretti, L. Villa, G. Vistoli, VEGA - an open platform to develop chemo-bio-informatics applications, using plug-in architecture and script programming, J. Comput. Aided. Mol. Des. 18 (2004) 167–173. doi:https://doi.org/10.1093/nar/gky1033.

[28] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, E.E. Bolton, PubChem 2019 update: improved access to chemical data, Nucleic Acids Res. 47 (2019) D1102–D1109. doi:https://doi.org/10.1093/nar/gky1033.

[29] I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V.A. Palyulin, E.V. Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, V.V. Prokopenko, Virtual computational chemistry laboratory – design and description, J. Comput. Aided. Mol. Des. 19 (2005) 453–463. doi:https://doi.org/10.1007/s10822-005-8694-y.

[30] C.F. Poole, The essence of chromatography, Elsevier, Amsterdam, The Netherlands, 2003.

[31] P. Broto, G. Moreau, C. Vandycke, Molecular structures: perception, autocorrelation descriptor and sar studies, Eur. J. Med. Chem. 19 (1984) 71–78.

[32] C.V. Eadsforth, P. Moser, Assessment of reverse-phase chromatographic methods for determining partition coefficients, Chemosphere. 12 (1983) 1459–1475. doi:https://doi.org/10.1016/0045-6535(83)90076-0.

[33] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, Weinheim, Germany, 2000.

[34] Talete, List of molecular descriptors calculated by Dragon, (n.d.). http://www.talete.mi.it/products/dragon_molecular_descriptor_list.pdf (accessed July 15, 2020).

Journal of Pharmaceutical and Biomedical Analysis

Supplementary material

Comparison of in-silico modelling and reversed-phase liquid chromatographic retention on an octadecyl silica column to predict skin permeability of pharmaceutical and cosmetic compounds

Yasmine Grooten, Aleksandra Sych, Debby Mangelings, Yvan Vander Heyden*

Vrije Universiteit Brussel (VUB), department of Analytical Chemistry, Applied Chemometrics and Molecular Modelling, Brussels, Belgium

*: Corresponding author. E-mail address: yvan.vander.heyden@vub.be. Vrije Universiteit Brussel (VUB), department of Analytical Chemistry, Applied Chemometrics and Molecular Modelling, Laarbeeklaan 103, B-1090 Brussels, Belgium.

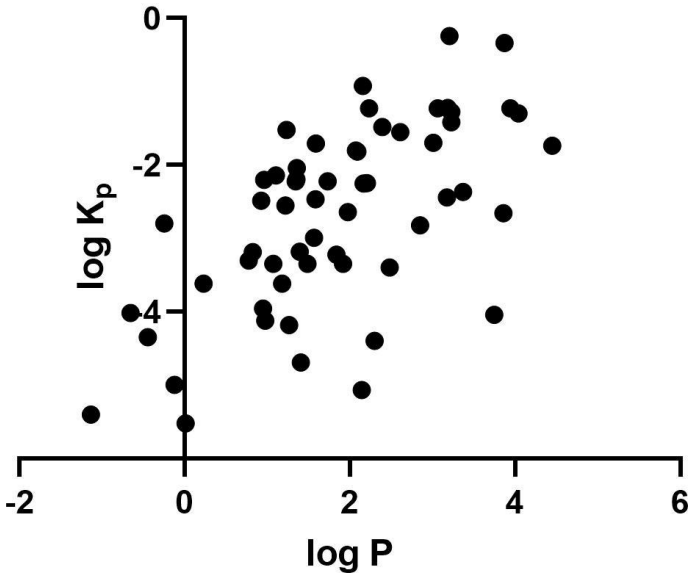

Fig. S1. log K_p versus log P values of the test set compounds.

Table S1. The experimental log *k* at the indicated ACN fraction, at pH 5.5. The value in bold was regarded as an outlier for the estimation of log k_w, the dashes indicate that no log *k* value was obtained.

Compound	log $k_{0.25}$	log $k_{0.30}$	log $k_{0.35}$	log $k_{0.40}$	log $k_{0.45}$
17α-Hydroxyprogesterone	1.696	1.359	0.987	0.725	0.589
2,4,6-Trichlorophenol	1.636	1.402	1.091	0.850	0.729
2,4-Dichlorophenol	1.449	1.236	0.957	0.728	0.639
2-Amino-4-nitrophenol	0.501	0.364	0.174	0.062	-0.032
2-Nitro-p-phenylenediamine	0.110	0.006	-0.129	-0.181	-0.257
4-Amino-2-nitrophenol	0.402	0.301	0.145	0.054	-0.025
Acetylsalicylic acid	-0.503	-0.517	-0.535	**-0.667**	-0.618
Aminopyrine	0.204	0.023	-0.163	-0.255	-0.304
Amylobarbital	0.915	0.692	0.461	0.271	0.163
Antipyrine	-0.081	-0.246	-0.434	-0.468	-0.486
Atropine	-0.216	-0.422	-0.515	-0.556	-0.605
Barbital	0.006	-0.129	-0.283	-0.339	-0.433
Benzoic acid	-0.503	-0.508	-0.546	-0.590	-0.644
Benzyl alcohol	0.225	0.108	-0.035	-0.113	-0.171
Caffeine	-0.421	-0.508	-0.515	-0.579	-0.581
Chloroxylenol	1.596	1.344	1.042	0.815	0.686
Chlorpheniramine (maleate)	0.585	0.292	-0.018	-0.214	-0.336
Cortexolone	1.148	0.851	0.518	0.278	0.177
Cortexone	1.533	1.203	0.827	0.559	0.433
Corticosterone	1.057	0.713	0.409	0.184	0.089
Cortisone	0.760	0.444	0.104	-0.041	-0.149
Diclofenac	1.278	0.862	0.427	0.123	0.006
Ephedrine(.HCl)	-0.503	-0.517	-0.535	-0.579	-0.605
Estriol	0.803	0.503	0.176	-0.041	-0.141
Estrone	1.674	1.451	1.107	0.853	0.701
Ethyl nicotinate	0.410	0.253	0.075	-0.018	-0.074
Flurbiprofen	1.083	0.728	0.334	0.085	-0.045
Haloperidol	1.218	0.878	0.502	0.260	0.127
Hydrocortisone	0.711	0.407	0.070	-0.072	-0.181
Ibuprofen	1.293	0.952	0.632	0.349	0.247
Indomethacin	1.368	0.931	0.490	0.160	0.041
Ketoprofen	0.566	0.288	-0.012	-0.266	-0.357
Lidocaine	0.243	0.184	0.077	0.029	0.044
m-Cresol	0.753	0.618	0.432	0.289	0.205
Methyl-4-hydroxybenzoate	0.654	0.495	0.300	0.149	0.060
Methyl nicotinate	0.075	-0.043	-0.172	-0.234	-0.298
m-Nitrophenol	0.849	0.701	0.501	0.341	0.243
Naproxen	0.717	0.401	0.116	-0.101	-0.205
o-Chlorophenol	0.900	0.745	0.554	0.392	0.303
o-Cresol	0.806	0.677	0.491	0.355	0.261
Paracetamol	-0.340	-0.498	-0.496	-0.567	-0.516
p-Cresol	0.764	0.614	0.429	0.286	0.207

Phenobarbitone	0.560	0.364	0.136	0.003	-0.089
Phenol	0.484	0.366	0.205	0.111	0.033
Piroxicam	0.449	-0.040	-0.086	-0.339	-0.450
p-Nitrophenol	0.796	0.638	0.411	0.283	0.187
p-Phenylenediamine	-	-	-	-	-
Prednisolone	0.688	0.380	0.032	-0.109	-0.220
Progesterone	1.981	1.637	1.296	1.130	0.896
Resorcinol	-0.045	-0.153	-0.283	-0.339	-0.417
Salicylic acid	-0.513	-0.508	-0.515	-0.614	-0.631
Testosterone	1.483	1.145	0.778	0.543	0.448
Thiourea	-	-	-	-	-
Thymol	1.758	1.387	1.198	0.972	0.827
Triamcinolone	0.400	0.085	-0.185	-0.373	-0.468
Triamcinolone acetonide	1.158	0.817	0.496	0.257	0.130
β-estradiol	1.729	1.388	1.007	0.740	0.615
β-Naphthol	1.280	1.053	0.801	0.587	0.473

Table S2. The extrapolated log *k* at a pure aqueous mobile phase, log k_w, the slope *s* and determination coefficient r^2, at pH 5.5, for eq. 3. The dash indicates that no log *k* value could be calculated.

Compound	**log $k_{w,pH5.5}$**	***s***	**r^2**
17α-Hydroxyprogesterone	3.064	0.057	0.974
z2,4,6-Trichlorophenol	2.797	0.047	0.982
2,4-Dichlorophenol	2.492	0.043	0.977
2-Amino-4-nitrophenol	1.171	0.027	0.985
2-Nitro-p-phenylenediamine	0.554	0.018	0.976
4-Amino-2-nitrophenol	0.946	0.022	0.986
Acetylsalicylic acid	-0.347	0.006	0.941
Aminopyrine	0.808	0.026	0.945
Amylobarbital	1.849	0.039	0.984
Antipyrine	0.380	0.021	0.871
Atropine	0.176	0.018	0.883
Barbital	0.526	0.022	0.973
Benzoic acid	-0.303	0.007	0.938
Benzyl alcohol	0.713	0.020	0.973
Caffeine	-0.247	0.008	0.893
Chloroxylenol	2.741	0.047	0.983
Chlorpheniramine (maleate)	1.706	0.047	0.972
Cortexolone	2.355	0.050	0.969
Cortexone	2.901	0.057	0.973
Corticosterone	2.216	0.049	0.961
Cortisone	1.835	0.046	0.948
Diclofenac	2.836	0.066	0.962
Ephedrine(.HCl)	-0.362	0.005	0.960
Estriol	1.962	0.049	0.966

Estrone	2.938	0.051	0.986
Ethyl nicotinate	0.996	0.025	0.959
Flurbiprofen	2.467	0.058	0.967
Haloperidol	2.557	0.056	0.970
Hydrocortisone	1.771	0.045	0.949
Ibuprofen	2.582	0.054	0.971
Indomethacin	2.995	0.068	0.963
Ketoprofen	1.723	0.048	0.973
Lidocaine	0.503	0.011	0.874
m-Cresol	1.457	0.029	0.986
Methyl-4-hydroxybenzoate	1.405	0.031	0.985
Methyl nicotinate	0.522	0.019	0.972
m-Nitrophenol	1.626	0.031	0.989
Naproxen	1.828	0.047	0.968
o-Chlorophenol	1.662	0.031	0.988
o-Cresol	1.506	0.028	0.989
Paracetamol	-0.188	0.008	0.610
p-cresol	1.469	0.029	0.984
Phenobarbitone	1.356	0.033	0.973
Phenol	1.050	0.023	0.984
Piroxicam	1.375	0.042	0.908
p-Nitrophenol	1.564	0.031	0.979
p-phenylenediamine	-	-	-
Prednisolone	1.768	0.046	0.948
Progesterone	3.262	0.054	0.981
Resorcinol	0.403	0.019	0.979
Salicylic acid	-0.316	0.007	0.789
Testosterone	2.748	0.053	0.959
Thiourea	-	-	-
Thymol	2.823	0.046	0.972
Triamcinolone	1.427	0.044	0.961
Triamcinolone acetonide	2.403	0.052	0.972
β-estradiol	3.109	0.058	0.971
β-Naphthol	2.294	0.042	0.985

Table S3. The experimental log *k* at the indicated ACN fraction, at pH 7. The values in bold were considered outliers for the determination of log k_w. The dash indicates that no log *k* value could be calculated.

Compound	log $k_{0.25}$	log $k_{0.30}$	log $k_{0.35}$	log $k_{0.40}$
17α-Hydroxyprogesterone	1.720	1.109	0.822	0.652
2,4,6-Trichlorophenol	0.553	0.388	0.201	0.000
2,4-Dichlorophenol	1.484	0.906	0.809	0.609
2-Amino-4-nitrophenol	-0.110	-0.098	-0.167	-0.332
2-Nitro-p-phenylenediamine	0.003	-0.073	-0.100	-0.209
4-Amino-2-nitrophenol	0.198	0.187	0.096	0.021
Acetylsalicylic acid	-0.501	-0.520	-0.526	-0.556
Aminopyrine	0.046	-0.056	-0.154	-0.250
Amylobarbital	0.573	0.425	0.326	0.129
Antipyrine	-0.252	-0.360	-0.378	-0.525
Atropine	-0.397	**-0.510**	-0.459	-0.515
Barbital	-0.209	-0.259	-0.292	-0.515
Benzoic acid	-	-	-	-
Benzyl alcohol	0.178	0.072	-0.003	-0.129
Caffeine	-0.626	-0.530	-0.477	-0.505
Chloroxylenol	1.321	1.133	0.901	0.671
Chlorpheniramine (maleate)	0.468	0.237	0.076	-0.079
Cortexolone	1.131	0.505	0.394	0.224
Cortexone	1.507	0.792	0.680	0.515
Corticosterone	0.615	0.505	0.287	0.120
Cortisone	0.389	0.288	0.068	-0.101
Diclofenac	0.434	0.161	-0.081	-0.214
Ephedrine(.HCl)	-0.474	-0.510	**-0.468**	-0.525
Estriol	0.401	0.175	0.071	-0.129
Estrone	1.840	1.245	0.958	0.704
Ethyl nicotinate	0.256	0.179	0.076	-0.044
Flurbiprofen	0.200	-0.069	-0.286	-0.515
Haloperidol	1.164	0.923	0.700	0.460
Hydrocortisone	0.333	0.225	0.027	-0.172
Ibuprofen	0.231	-0.009	-0.235	-0.505
Indomethacin	0.497	0.214	-0.100	-0.556
Ketoprofen	0.373	**-0.571**	-0.425	-0.515
Lidocaine	0.942	0.772	0.681	0.555
m-Cresol	0.548	0.476	0.385	0.236
Methyl-4-hydroxybenzoate	0.396	0.313	0.231	0.070
Methyl Nicotinate	-0.052	-0.098	-0.141	-0.295
m-Nitrophenol	0.626	0.505	0.421	0.298
Naproxen	-0.232	-0.402	-0.425	-0.525
o-Chlorophenol	0.669	0.542	0.481	0.314
o-Cresol	0.607	0.539	0.437	0.246
Paracetamol	-0.492	-0.520	**-0.477**	-0.515

p-cresol	0.549	0.481	0.377	0.203
Phenobarbitone	0.168	0.128	0.003	-0.133
Phenol	0.331	0.298	0.199	0.120
Piroxicam	0.333	-0.367	-0.401	-0.525
p-Nitrophenol	0.186	0.157	0.060	-0.113
p-phenylenediamine	-	-	-	-
Prednisolone	0.298	0.196	-0.006	-0.129
Progesterone	1.838	1.524	1.291	0.872
Resorcinol	-0.161	-0.200	-0.246	-0.505
Salicylic acid	-0.580	-0.530	-0.496	-0.515
Testosterone	1.132	0.895	0.651	0.483
Thiourea	-	-	-1.450	-1.681
Thymol	1.708	1.419	1.198	0.940
Triamcinolone	0.085	-0.066	-0.230	-0.410
Triamcinolone acetonide	1.160	0.496	0.363	0.151
β-estradiol	1.339	1.066	0.827	0.560
β-Naphthol	1.306	0.781	0.696	0.584

Table S4. The extrapolated log *k* at a pure aqueous mobile phase, log k_w, the slope *s* and determination coefficient r^2 at pH 7, for eq. 3. The dash indicates that no values could be calculated.

Compound	log $k_{w,pH7}$	*s*	r^2
17*α*-Hydroxyprogesterone	3.345	0.070	0.923
2,4,6-Trichlorophenol	1.485	0.037	0.998
2,4-Dichlorophenol	2.721	0.054	0.875
2-Amino-4-nitrophenol	0.302	0.015	0.774
2-Nitro-p-phenylenediamine	0.336	0.013	0.951
4-Amino-2-nitrophenol	0.531	0.012	0.930
Acetylsalicylic acid	-0.553	0.003	0.937
Aminopyrine	0.537	0.020	1.000
Amylobarbital	1.291	0.029	0.983
Antipyrine	0.164	0.017	0.927
Atropine	-0.202	0.008	0.975
Barbital	0.299	0.019	0.825
Benzoic acid	-	-	-
Benzyl alcohol	0.677	0.020	0.991
Caffeine	-0.472	-0.008	0.687
Chloroxylenol	2.425	0.044	0.998
Chlorpheniramine (maleate)	1.348	0.036	0.990
Cortexolone	-0.334	0.057	0.855
Cortexone	2.882	0.062	0.831
Corticosterone	1.489	0.034	0.986
Cortisone	1.259	0.034	0.982
Diclofenac	1.496	0.044	0.979
Ephedrine(.HCl)	-0.405	0.003	0.824

Estriol	1.231	0.034	0.983
Estrone	3.589	0.074	0.954
Ethyl nicotinate	0.769	0.020	0.991
Flurbiprofen	1.368	0.047	0.998
Haloperidol	2.330	0.047	1.000
Hydrocortisone	1.215	0.034	0.984
Ibuprofen	1.454	0.049	0.999
Indomethacin	2.271	0.069	0.987
Ketoprofen	1.883	0.062	0.943
Lidocaine	1.550	0.025	0.986
m-Cresol	1.079	0.021	0.972
Methyl-4-hydroxybenzoate	0.941	0.021	0.968
Methyl nicotinate	0.357	0.015	0.893
m-Nitrophenol	1.156	0.021	0.995
Naproxen	0.189	0.018	0.915
o-Chlorophenol	1.234	0.023	0.971
o-Cresol	1.226	0.024	0.947
Paracetamol	-0.469	0.001	0.404
p-cresol	1.144	0.023	0.958
Phenobarbitone	0.711	0.021	0.954
Phenol	0.713	0.015	0.968
Piroxicam	1.454	0.052	0.753
p-Nitrophenol	0.718	0.020	0.903
p-phenylenediamine	-	-	-
Prednisolone	1.054	0.030	0.985
Progesterone	3.416	0.063	0.987
Resorcinol	0.424	0.022	0.804
Salicylic acid	-0.679	-0.005	0.675
Testosterone	2.214	0.044	0.994
Thiourea	0.167	-0.046	1.000
Thymol	2.957	0.050	0.998
Triamcinolone	0.918	0.033	0.998
Triamcinolone acetonide	2.597	0.063	0.877
β-estradiol	2.622	0.052	0.999
β-Naphthol	2.304	0.045	0.825

Table S5. The best stepwise MLR models including the retention factors log *k* and the Vega ZZ descriptors (when excluding the virtual log *P*).

RMSEC	RMSECV	r^2	n	Equation
0.739	0.806	0.65	56	$\log K_p = -1.51 + 0.68 \log k_{0.35.\ pH\ 5.5} - 0.044$ Atoms - 0.39 HbDon + 0.30 Lipole (Broto)
0.740	0.807	0.65	56	$\log K_p = -1.53 + 0.60 \log k_{0.30.\ pH\ 5.5} - 0.046$ Atoms - 0.39 HbDon + 0.29 Lipole (Broto)
0.732	0.800	0.66	56	$\log K_p = -1.54 + 0.58 \log k_{0.25.\ pH\ 5.5} - 0.049$ Atoms - 0.39 HbDon + 0.28 Lipole (Broto)
0.722	0.792	0.67	56	$\log K_p = -1.61 + 0.42 \log k_{w.\ pH\ 5.5} - 0.053$ Atoms - 0.39 HbDon + 0.25 Lipole (Broto)
0.741	0.811	0.65	56	$\log K_p = -1.72 + 0.33 \log k_{w.\ pH\ 7} - 0.39$ Atoms - 0.047 HbDon + 0.33 Lipole (Broto)

Table S6. The best automatic MLR models with the retention factors from 45% v/v ACN at pH 5.5 and the Vega ZZ descriptors (n = 56).

Nr	RMSEC	RMSECV	r^2	Equation
1	1.183	1.223	0.10	$\log K_p = -2.65 + 0.96 \log k_{0.45,\ pH\ 5.5}$
2	0.884	0.920	0.50	$\log K_p = -1.02 + 1.16 \log k_{0.45,\ pH\ 5.5} - 0.053$ Atoms
3	0.727	0.783	0.66	$\log K_p = -2.19 + 0.35 \log k_{0.45,\ pH\ 5.5} - 0.046$ Atoms + 0.51 Virtual log P
4	0.710	0.778	0.68	$\log K_p = -1.88 + 0.35 \log k_{0.45,\ pH\ 5.5} - 0.044$ Atoms + 0.45 Virtual log P - 0.18 HbDon
5	0.693	0.779	0.69	$\log K_p = -2.08 + 0.30 \log k_{0.45,\ pH\ 5.5} - 0.041$ Atoms + 0.37 Virtual log P - 0.24 HbDon + 0.17 Lipole (Broto)

Table S7. The best MLR models with the retention factors from 40% v/v ACN at pH 5.5 and the Vega ZZ descriptors (n = 56).

Nr	RMSEC	RMSECV	r^2	Equation
1	1.192	1.231	0.09	$\log K_p = -2.72 + 0.83 \log k_{0.40,\ pH\ 5.5}$
2	0.884	0.920	0.50	$\log K_p = -1.08 + 1.08 \log k_{0.40,\ pH\ 5.5} - 0.054$ Atoms
3	0.728	0.784	0.66	$\log K_p = -2.21 + 0.31 \log k_{0.40,\ pH\ 5.5} - 0.046$ Atoms + 0.51 Virtual log P
4	0.711	0.778	0.68	$\log K_p = -1.90 + 0.31 \log k_{0.40,\ pH\ 5.5} - 0.044$ Atoms + 0.45 Virtual log P - 0.19 HbDon
5	0.694	0.779	0.69	$\log K_p = -2.10 + 0.27 \log k_{0.40,\ pH\ 5.5} - 0.042$ Atoms + 0.36 Virtual log P - 0.24 HbDon + 0.18 Lipole (Broto)

Table S8. The best MLR models with the retention factors from 35% v/v ACN at pH 5.5 and the Vega ZZ descriptors (*n* = 56).

Nr	RMSEC	RMSECV	r^2	Equation
1	1.190	1.229	0.09	$\log K_p = -2.84 + 0.76 \log k_{0.35,\ pH\ 5.5}$
2	0.853	0.889	0.53	$\log K_p = -1.18 + 1.08 \log k_{0.35,\ pH\ 5.5} - 0.056$ Atoms
3	0.722	0.779	0.67	$\log K_p = -2.18 + 0.38 \log k_{0.35,\ pH\ 5.5} - 0.048$ Atoms + 0.48 Virtual log P
4	0.704	0.771	0.68	$\log K_p = -1.85 + 0.39 \log k_{0.35,\ pH\ 5.5} - 0.046$ Atoms + 0.41 Virtual log P - 0.19 HbDon
5	0.688	0.773	0.70	$\log K_p = -2.05 + 0.35 \log k_{0.35,\ pH\ 5.5} - 0.043$ Atoms + 0.34 Virtual log P - 0.24 HbDon + 0.17 Lipole (Broto)

Table S9. The best MLR models with the retention factors from 30% v/v ACN at pH 5.5 and the Vega ZZ descriptors (*n* = 56).

Nr	RMSEC	RMSECV	r^2	Equation
1	1.210	1.249	0.06	$\log K_p = -2.90 + 0.52 \log k_{0.30,\ pH\ 5.5}$
2	0.850	0.888	0.54	$\log K_p = -1.26 + 0.95 \log k_{0.30,\ pH\ 5.5} - 0.060$ Atoms
3	0.724	0.783	0.66	$\log K_p = -2.21 + 0.31 \log k_{0.30,\ pH\ 5.5} - 0.049$ Atoms + 0.48 Virtual log P
4	0.705	0.774	0.68	$\log K_p = -1.87 + 0.34 \log k_{0.30,\ pH\ 5.5} - 0.047$ Atoms + 0.41 Virtual log P - 0.20 HbDon
5	0.690	0.777	0.69	$\log K_p = -2.06 + 0.29 \log k_{0.30,\ pH\ 5.5} - 0.044$ Atoms + 0.34 Virtual log P - 0.25 HbDon + 0.17 Lipole (Broto)

Table S10. The best MLR models with the retention factors from 25% v/v ACN at pH 5.5 and the Vega ZZ descriptors (*n* = 56).

Nr	RMSEC	RMSECV	r^2	Equation
1	1.224	1.263	0.04	$\log K_p = -2.90 + 0.37 \log k_{0.25,\ pH\ 5.5}$
2	0.840	0.880	0.55	$\log K_p = -1.30 + 0.89 \log k_{0.25,\ pH\ 5.5} - 0.064$ Atoms
3	0.723	0.783	0.66	$\log K_p = -2.21 + 0.30 \log k_{0.25,\ pH\ 5.5} - 0.050$ Atoms + 0.47 Virtual log P
4	0.703	0.773	0.68	$\log K_p = -1.85 + 0.33 \log k_{0.25,\ pH\ 5.5} - 0.049$ Atoms + 0.40 Virtual log P - 0.20 HbDon
5	0.688	0.776	0.70	$\log K_p = -2.05 + 0.29 \log k_{0.25,\ pH\ 5.5} - 0.045$ Atoms + 0.32 Virtual log P - 0.25 HbDon + 0.17 Lipole (Broto)

Table S11. The best MLR models with the extrapolated retention factors (log k_w) at pH 5.5 and the Vega ZZ descriptors (n = 56).

Nr	RMSEC	RMSECV	r^2	Equation
1	1.236	1.276	0.02	$\log K_p = -2.90 + 0.16 \log k_{w,\ pH\ 5.5}$
2	0.826	0.871	0.56	$\log K_p = -1.49 + 0.62 \log k_{w,\ pH\ 5.5}$ - 0.069 Atoms
3	0.722	0.787	0.66	$\log K_p = -2.25 + 0.22 \log k_{w,\ pH\ 5.5}$ - 0.052 Atoms + 0.46 Virtual log P
4	0.699	0.774	0.69	$\log K_p = -1.86 + 0.26 \log k_{w,\ pH\ 5.5}$ - 0.051 Atoms + 0.37 Virtual log P - 0.22 HbDon
5	0.685	0.777	0.70	$\log K_p = -2.05 + 0.23 \log k_{w,\ pH\ 5.5}$ - 0.048 Atoms + 0.30 Virtual log P - 0.26 HbDon + 0.16 Lipole (Broto)

Table S12. The best MLR models with the retention factors from 40% v/v ACN at pH 7 and the Vega ZZ descriptors (n = 56).

Nr	RMSEC	RMSECV	r^2	Equation
1	1.229	1.269	0.04	$\log K_p = -2.67 + 0.49 \log k_{0.40,\ pH\ 7}$
2	0.962	1.011	0.41	$\log K_p = -1.29 + 0.30 \log k_{0.40,\ pH\ 7}$ - 0.010 Melting point
3	0.729	0.788	0.66	$\log K_p = -1.72 - 0.055 \log k_{0.40,\ pH\ 7}$ - 0.81 Gyration radius + 0.74 Virtual log P
4	0.702	0.777	0.69	$\log K_p = -1.71 - 0.097 \log k_{0.40,\ pH\ 7}$ - 0.58 Gyration radius + 0.65 Virtual log P - 0.0038 Melting point
5	0.697	0.786	0.69	$\log K_p = -2.28 + 0.048 \log k_{0.40,\ pH\ 7}$ - 0.039 Atoms + 0.43 Virtual log P - 0.24 HbDon + 0.18 Lipole (Broto)

Table S13. The best MLR models with the retention factors from 35% v/v ACN at pH 7 and the Vega ZZ descriptors (n = 56).

Nr	RMSEC	RMSECV	r^2	Equation
1	1.229	1.267	0.04	$\log K_p = -2.75 + 0.46 \log k_{0.35,\ pH\ 7}$
2	0.959	1.007	0.41	$\log K_p = -1.34 + 0.31 \log k_{0.35,\ pH\ 7}$ - 0.010 Melting point
3	0.729	0.787	0.66	$\log K_p = -1.73 - 0.082 \log k_{0.35,\ pH\ 7}$ - 0.80 Gyration radius + 0.74 Virtual log P
4	0.701	0.777	0.69	$\log K_p = -1.71 - 0.11 \log k_{0.35,\ pH\ 7}$ - 0.58 Gyration radius + 0.65 Virtual log P - 0.0038 Melting point
5	0.697	0.786	0.69	$\log K_p = -2.29 + 0.048 \log k_{0.35,\ pH\ 7}$ - 0.039 Atoms + 0.42 Virtual log P - 0.24 HbDon + 0.18 Lipole (Broto)

Table S14. The best MLR models with the retention factors from 30% v/v ACN at pH 7 and the Vega ZZ descriptors (n = 55).

Nr	RMSEC	RMSECV	r^2	Equation
1	1.239	1.280	0.02	$\log K_p = -2.75 + 0.30 \log k_{0.30,\ pH\ 7}$
2	0.945	0.988	0.43	$\log K_p = -1.12 + 0.71 \log k_{0.30,\ pH\ 7} - 0.055$ Atoms
3	0.726	0.784	0.66	$\log K_p = -1.60 - 0.14 \log k_{0.30,\ pH\ 7} - 0.83$ Gyration radius + 0.73 Virtual log P
4	0.703	0.783	0.68	$\log K_p = -1.61 - 0.13 \log k_{0.30,\ pH\ 7} - 0.62$ Gyration radius + 0.65 Virtual log P - 0.0034 Melting point
5	0.703	0.798	0.68	$\log K_p = -2.27 + 0.049 \log k_{0.30,\ pH\ 7} - 0.040$ Atoms + 0.42 Virtual log P - 0.23 HbDon + 0.17 Lipole (Broto)

Table S15. The best MLR models with the retention factors from 25% v/v ACN at pH 7 and the Vega ZZ descriptors (n = 55).

Nr	RMSEC	RMSECV	r^2	Equation
1	1.239	1.278	0.02	$\log K_p = -2.79 + 0.25 \log k_{0.25,\ pH\ 7}$
2	0.901	0.938	0.48	$\log K_p = -1.12 + 0.77 \log k_{0.25,\ pH\ 7} - 0.061$ Atoms
3	0.729	0.784	0.66	$\log K_p = -1.56 - 0.014 \log k_{0.25,\ pH\ 7} - 0.84$ Gyration radius + 0.71 Virtual log P
4	0.706	0.783	0.68	$\log K_p = -1.56 + 0.0057 \log k_{0.25,\ pH\ 7} - 0.63$ Gyration radius + 0.63 Virtual log P - 0.0035 Melting point
5	0.697	0.788	0.69	$\log K_p = -2.19 + 0.17 \log k_{0.25,\ pH\ 7} - 0.043$ Atoms + 0.39 Virtual log P - 0.22 HbDon + 0.16 Lipole (Broto)

Table S16. The best MLR models with the extrapolated retention factors (log k_w) at pH 7 and the Vega ZZ descriptors (n = 56).

Nr	RMSEC	RMSECV	r^2	Equation
1	1.229	1.268	0.04	$\log K_p = -2.95 + 0.22 \log k_{w,\ pH\ 7}$
2	0.897	0.940	0.49	$\log K_p = -1.50 + 0.57 \log k_{w,\ pH\ 7} - 0.060$ Atoms
3	0.724	0.782	0.67	$\log K_p = -2.44 + 0.17 \log k_{w,\ pH\ 7} - 0.047$ Atoms + 0.53 Virtual log P
4	0.703	0.770	0.68	$\log K_p = -2.06 + 0.17 \log k_{w,\ pH\ 7} - 0.046$ Atoms + 0.45 Virtual log P - 0.20 HbDon
5	0.685	0.769	0.70	$\log K_p = -2.21 + 0.17 \log k_{w,\ pH\ 7} - 0.044$ Atoms + 0.35 Virtual log P - 0.24 HbDon + 0.18 Lipole (Broto)

Table S17. The best PLS models combining retention factors with the E-Dragon descriptors.

	% ACN	PLS factors	RMSEC	RMSECV	r^2	n
pH 5.5	45	6	0.670	0.888	0.72	56
	40	6	0.670	0.888	0.72	56
	35	6	0.670	0.888	0.72	56
	30	6	0.670	0.888	0.72	56
	25	6	0.670	0.888	0.72	56
	0	6	0.670	0.887	0.72	56
pH 7	40	6	0.672	0.871	0.72	56
	35	6	0.672	0.871	0.72	56
	30	6	0.675	0.888	0.72	55
	25	6	0.675	0.888	0.72	55
	0	6	0.672	0.870	0.72	56

Table S18. The best PLS models combining retention factors with the Vega ZZ descriptors.

	% ACN	PLS factors	RMSEC	RMSECV	r^2	n
pH 5.5	45	5	0.736	0.832	0.65	56
	40	6	0.699	0.811	0.69	56
	35	5	0.732	0.862	0.66	56
	30	5	0.732	0.832	0.66	56
	25	6	0.695	0.825	0.69	56
	0	5	0.725	0.803	0.66	56
pH 7	40	5	0.760	0.825	0.63	56
	35	5	0.760	0.809	0.63	56
	30	5	0.740	0.861	0.65	55
	25	6	0.698	0.834	0.69	55
	0	6	0.701	0.818	0.69	56