Chemistry Europe
European Chemical Societies Publishing

WILEY VCH

Excellence in Chemistry Research

Announcing our new flagship journal

- Gold Open Access
- Publishing charges waived
- Preprints welcome
- Edited by active scientists

Meet the Editors of *ChemistryEurope*

Luisa De Cola
Università degli Studi di Milano Statale, Italy

Ive Hermans
University of Wisconsin-Madison, USA

Ken Tanaka
Tokyo Institute of Technology, Japan

ChemPhysChem

Chemistry Europe
European Chemical Societies Publishing

Accepted Article

Title: Critical Assessment of pH-Dependent Lipophilicity Profiles of Small Molecules: Which One Should We Use and In Which Cases?

Authors: Esteban Bertsch, Sebastián Suñer, Silvana Pinheiro, and William J. Zamora

To be cited as: *ChemPhysChem* **2023**, e202300548

Link to VoR: https://doi.org/10.1002/cphc.202300548

WILEY VCH

10.1002/cphc.202300548

Critical Assessment of pH-Dependent Lipophilicity Profiles of Small Molecules: Which One Should We Use and In Which Cases?

Esteban Bertsch[1+], Sebastián Suñer[1+], Dr. Silvana Pinheiro[1,2], Prof. Dr. William J. Zamora[1,2,3,*]

1. CBio[3] Laboratory, School of Chemistry, University of Costa Rica, San Pedro, San José, Costa Rica
2. Laboratory of Computational Toxicology and Artificial Intelligence (LaToxCIA), Biological Testing Laboratory (LEBi), University of Costa Rica, San Pedro, San José, Costa Rica.
3. Advanced Computing Lab (CNCA), National High Technology Center (CeNAT), Pavas, San José, Costa Rica

* Corresponding author: william.zamoraramirez@ucr.ac.cr

[+]These authors contributed equally to this work

Keywords Partition coefficient, Lipophilicity profiles, Machine learning, Chemoinformatics, Drug Design, Ion pair partitioning.

10.1002/cphc.202300548

Abstract

Lipophilicity is a physicochemical property with wide relevance in drug design, computational biology, food, environmental and medicinal chemistry. Lipophilicity is commonly expressed as the partition coefficient for neutral molecules, whereas for molecules with ionizable groups, the distribution coefficient (D) at a given pH is used. The logD_{pH} is usually predicted using a pH correction over the logP_N using the pK_a of ionizable molecules, while often ignoring the apparent ion pair partitioning (P_{IP}^{app}). In this work, we studied the impact of P_{IP}^{app} on the prediction of both the experimental lipophilicity of small molecules and experimental lipophilicity-based applications and metrics such as lipophilic efficiency (LipE), distribution of spiked drugs in milk products, and pH-dependent partition of water contaminants in synthetic passive samples such as silicones. Our findings show that better predictions are obtained by considering the apparent ion pair partitioning. In this context, we developed machine learning algorithms to determine the cases that P_I^{app} should be considered. The results indicate that small, rigid, and unsaturated molecules with logP_N close to zero, which present a significant proportion of ionic species in the aqueous phase, were better modeled using the apparent ion pair partitioning (P_{IP}^{app}). Finally, our findings can serve as guidance to the scientific community working in early-stage drug design, food, and environmental chemistry.

Introduction

Lipophilicity has been a relevant physicochemical property in pharmaceutical research since the late 1800s, where the toxicity and anesthetic properties of several substances have been correlated to their solubilities in water and oil/water partition coefficients.[1] In addition, this property has also been associated with several pharmacokinetic properties, such as enzyme binding[2], toxicity[3], solubility[4], membrane permeability[5], and bioaccumulation.[6] Thus, lipophilicity has been considered a significant descriptor in drug discovery metrics, such as Lipinski's[7] and Veber's[8] empirical rules, which are intended to optimize oral bioavailability for drug-like compounds. The partition coefficient (P_N) describes the equilibrium of a molecule between the organic and aqueous phases, where the *n*-octanol/water system has historically been the medium of choice in pharmaceutical research because of its high correlation with biological activities.[9,10] However, logP_N only describes the equilibrium of molecules in their neutral states, which implies an unrealistic protonation state for most molecules with ionizable groups at physiological pH.

Since the pH of the solution directly affects the concentration of neutral and ionic species, the equilibrium constant varies with pH, which also means that the lipophilicity of a compound is dependent on it. The partition coefficient as a function of pH is often called distribution coefficient (logD_{pH}).[11] The logD_{pH} is often considered to be a more proper descriptor than logP_N for human bioavailability due to the frequent pH-dependence of drugs. This property has been shown to be useful in QSAR models to explain how small molecules have human brain cell permeability[12] or bind to human serum albumin[13]. The logD_{pH} has also been used as an effective predictor of pH-dependent lipophilicity profiles for small molecules[14] and to characterize structural properties in proteins and peptides, such as protein-folding and aggregation[15], solubility[16], and antimicrobial activity[17,18], through pH-dependent lipophilicity scales.[19,20]

As an alternative to experimentally determined logD_{pH} values, theoretical lipophilicity profiles provide the opportunity to obtain this descriptor quickly and often with high accuracy.[14,21,22] Equation 1 models logD_{pH} as a function of pH for monoacidic and monobasic compounds. This equation is derived as the mass balance between the ionic and neutral species in

thermodynamic equilibrium in the aqueous phase. This model assumes that the organic phase holds mostly neutral species, so that the acid-base dissociation is negligible, and it also assumes that there is not a partition equilibrium for the ionic species (e.g., ion-pairs, counterions, molecular aggregates).[23]

$$\log D_{pH} = \log P_N - \log \left(1 + 10^{\delta}\right) \quad [1]$$

where $\delta = pH - pK_a$ for acids, and $\delta = pK_a - pH$ for bases.

Figure 1a shows the equilibria from which Eq. 1 is derived. This formalism has been used to easily calculate logD_{pH} from logP_N values obtained by empirical computational models.[24–26] This equation was widely used in logD_{pH} estimation methods in the SAMPL6 and SAMPL7 blind challenges, which is a large-scale comparative evaluation for drug design predictive models.[27,28]

10.1002/cphc.202300548

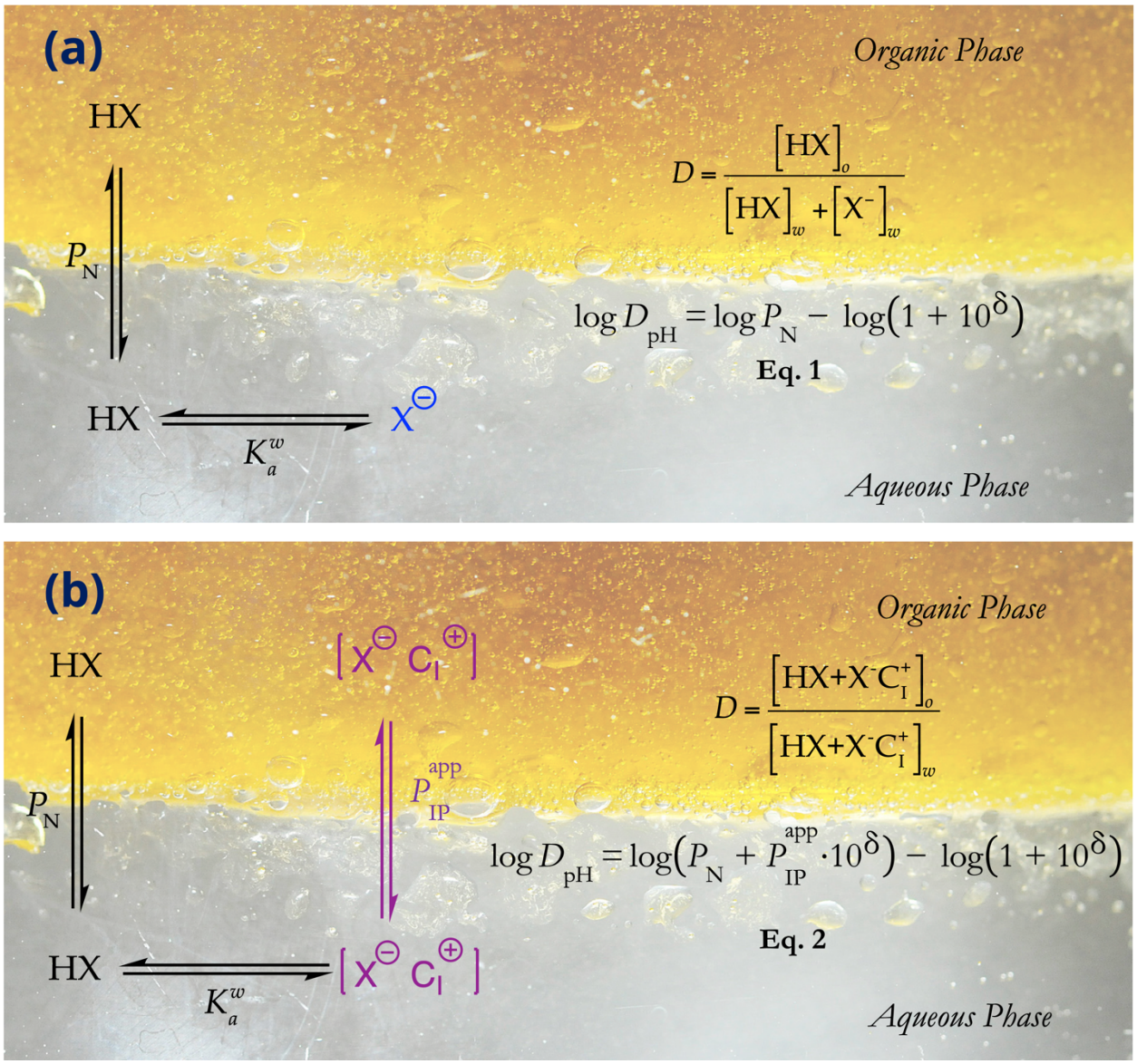


Figure 1. Representations of the partition mechanism for a symbolic ionizable acidic molecule for both neutral (**HX**) and ionic ($\mathbf{X^-}$, $\mathbf{C_I^+}$) species using **(a)** Equation 1 and **(b)** Equation 2. The theoretical partition of the charged species ($\mathbf{X^-}$, $\mathbf{C_I^+}$) was replaced by experimentally measurable apparent ion pair partitioning (P_{IP}^{app}) in Eq. 2.

Equation 2 represents the extended lipophilicity profile of monoprotic acids and bases (Fig. 1b). This model considers acid-base ionization in both water and *n*-octanol phases, where ionic species migrate between the phases

$$\log D_{\mathrm{pH}} = \log\left(P_{\mathrm{N}} + P_{\mathrm{IP}}^{\mathrm{app}} \cdot 10^{\delta}\right) - \log\left(1 + 10^{\delta}\right) \qquad [2]$$

Equation 2 is commonly called the ion pair partitioning ($P_{\mathrm{IP}}^{\mathrm{app}}$) model[29], which represents a simplification that considers only the partition of the charged organic species (see Figure 1b). Experimental techniques for lipophilicity evaluation such as shake-flask, potentiometric, and chromatographic methods[30], can measure but do not allow direct identification of the nature of the ionic species involved in the partitioning; hence, the partition of ionic species is measured as an apparent partitioning, usually assigned to ion pair partitioning ($P_{\mathrm{IP}}^{\mathrm{app}}$). This experimentally measurable apparent partition coefficient depends on the background salt[31] and compound concentration[32], and may involve many more complex species, such as ion-pairs[33–40] and aggregates[41]. Some studies have simplified the $P_{\mathrm{IP}}^{\mathrm{app}}$ to the partition of only ionic organic species (P_{I}) because these methods have been parametrized using experimental data of partitions to extreme pH without considering the corresponding counter-ions ($\mathbf{C_I}^+$ and $\mathbf{C_I}^-$), e.g. sodium or potassium ions for organic anions and chloride ions for organic cations.[14,42] However, theoretical studies have modeled it using the participation of ion-pairs (P_{IP})[21,22]. Recently, an alternative model[14] to ion-pair partitioning has been used by applying the theory of ionic transfer between two immiscible electrolyte solutions (ITIES)[43,44], obtaining excellent predictions of experimental logD_{pH} values. Previous experimental trials have also shown the importance of the $P_{\mathrm{IP}}^{\mathrm{app}}$ of ionizable molecules in *n*-octanol/water systems[33-40]. Recently, Disdier *et al*. measured the logD_{pH} at different pH values of a set of 13 compounds via the shake-flask method[45], where fitted experimental values to lipophilicity formalisms for mono- and poly substituted acids, amphoteric, and zwitterionic species derived on previous theoretical studies.[46] The relevance of $P_{\mathrm{IP}}^{\mathrm{app}}$ for small ionic molecules between aqueous and organic phases has also been studied through interphase transfer mechanisms of substances via ionic partition diagrams as a function of pH obtained through cyclic voltammetry.[47–49]

Despite the lack of a consensus formalism to model logD_{pH} as a function of P_{IP}^{app}, and considering that different theoretical approaches have shown similar trends[14,21,22], Equation 2 has been successfully used for modeling the lipophilicity of ionized compounds in many areas of basic and applied sciences. For instance, to study the aggregation of naphthenic acids in aqueous environments with different saline concentrations[50], in logD_{pH} calculations for lignin derivatives and small datasets of drug-like compounds in different solvents by QM and statistical thermodynamical methods[51] , partitioning of antioxidants[52], aquatic hazard assessment of ionizable organic chemicals[53], sorption mechanisms of antimicrobials in the soil[54], and physicochemical properties of peptides and proteins.[15–18]

A previous study has evaluated predictions of logD_{pH} using Equations 1 and 2 for a small set of 35 ionizable molecules with computed logP_N and logP_I^{app} values calculated via an extension of the Miertus-Scrocco-Tomassi solvation model.[14] In that work, Equation 1 tends to overestimate the hydrophobicity of the studied molecules, given that the P_{IP}^{app} is not considered, whereas Equation 2 predicts a logD_{pH} value closer to the experimental values. Thus, Equation 2 provided a more exact lipophilicity profile for those 35 ionizable molecules over a wider pH range than Equation 1.[14] However, no systematized study has been performed to evaluate the importance of considering the ion pair partitioning on the logD_{pH} prediction for large sets of small drug-like molecules at various pH values, although it has been reported that much of the poor performance of some models on blind challenges has been due to the simplification of ignoring the ionic species partition.[27]

In this study, our aim is to evaluate the impact of considering the P_{IP}^{app} in determining pH-dependent lipophilicity profiles of small molecules. We also aim to provide guidance to the scientific community working in early-stage drug design, food, and environmental chemistry, specifically those dealing with ionizable molecules. Our goal is to help researchers determine a priori which pH-dependent lipophilicity profile should be used based solely on structural features of the substance of interest. To this end, we collected the experimental values of log$P_{N,}$ pK_a, and logP_{IP}^{app} of different compounds at various pH values as well as experimental data of lipophilicity-based applications and metrics such as lipophilic efficiency (LipE), distribution of spiked drugs in

milk products and pH-dependent partition in passive samples, which were used to compute logD_{pH} with Equations 1 and 2. The predictions using both equations were then used to compare their performances using statistical parameters. Finally, logistic regression (**LR**), random forest classification (**RFC**), and support vector machine (**SVM**) models were developed to define from the molecular structure which formalism is recommended for modeling pH-dependent lipophilicity profiles.

Methodology

Data collection and classification

We critically compiled the experimental values of log$P_{N,}$ pK_a, logP_{IP}^{app}, and logD_{pH} of 225 entries based on earlier literature reports (database available in reference 33).[29,55,56] Refs. 29 and 55 were chosen based on the wide selection of experimental data for log$P_{N,}$ logD_{pH}, and logP_{IP}^{app} values and because they encompass the desired chemical space of small molecules for our modeling. SMILES codes were collected from publicly available data in PubChem.[57] The experimental pK_a values were also obtained from PubChem, but they were corroborated by reviewing their values in primary literature reports.[38,57–80] The experimental technique of logP_N, logD_{pH}, and logP_{IP}^{app}measurements for each entry were thoroughly revised and added to the database.[74,81–90] Ref 55 provided experimental logD_{pH} values of molecules in diverse pH ranges. The logP_{IP}^{app} values were obtained from the logD_{pH} at the most extreme measured pH, in which the molecule would be mostly (above 95 %) in its ionized state. The logP_{IP}^{app}values for molecules that were not measured under ionizable pH conditions were obtained from external sources.[38,74,91,92] The molecules were classified as acids or bases based on their functional groups and experimental pK_a values deposited in our database. Zwitterionic compounds were found by evaluating the difference between acidic and basic experimental pK_a in conjunction with ChemAxon's calculator of ionic species distribution as a function of pH.[93] Some amphoteric species were also classified as acidic or basic based on the behavior of their lipophilicity profiles, which were evaluated using the ChemAxon partitioning calculator.[94]

Figure 2 shows the distribution of the molecules along several descriptors of their chemical space. Most compounds can be considered small molecules because they tend to have small molecular weights (< 400 $gmol^{-1}$) and topological polar surface areas (< 100 $Å^2$). Our database consists mostly of lipophilic species since the clog*P* values are mostly positive, which coincides with the low polarity of our molecules, demonstrated by the tendency of low counts of hydrogen bond donors (< 3) and acceptors (< 7).

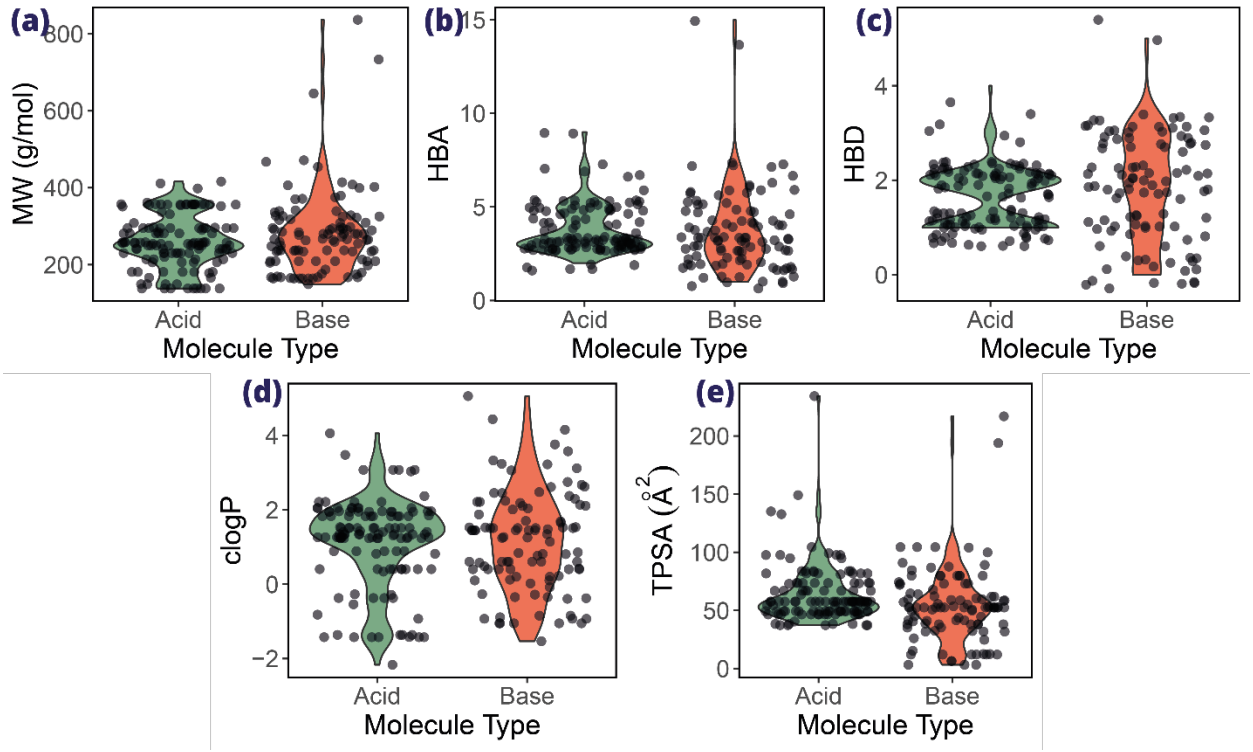


Figure 2. Distribution of molecular properties in the database[56] by (a) Molecular weight (MW), (b) hydrogen bond acceptors, (c) hydrogen bond donors, (d) calculated logP_N (obtained with Alogp)[95], and (d) topological polar surface area. These descriptors were calculated using the '*RCDK*' package in R.

Performance of pH-dependent lipophilicity profiles

The experimental data for each molecule were used to compute the logD_{pH} values using Eq. 1 and Eq. 2 and were labeled as log$D_{Eq.1}$ and log$D_{Eq.2}$, respectively. The modeling performance for each molecule was evaluated by calculating the absolute errors d_1 and d_2 (Eqs. 3 and 4):

$$d_1 = \left| \log D_{Eq.1} - \log D_{exp} \right| \quad [3]$$

$$d_2 = \left| \log D_{Eq.2} - \log D_{exp} \right| \quad [4]$$

where logD_{exp} represents the experimental logD_{pH} value.

The performance of the two formalisms was tested by performing a linear regression of log$D_{Eq.1}$ and log$D_{Eq.2}$ on their experimental values. The root-mean-squared error (RMSE), mean absolute error (MAE), mean squared error (MSE), and Pearson's correlation coefficient squared (R^2) were calculated with the '*Metrics*' package in R.[96] We also tested the performance of each formalism on each individual molecule using the descriptor $\boldsymbol{d_3}$ (Eq. 5). When d_3 yielded a value greater than zero, Eq. 2 fits better than Eq. 1. Accordingly, we created a binomial conditional based on the values of d_3, where Eq. 2 should be used when d_3 is greater than 0.2 (see Results and Discussion); otherwise, both equations are considered equivalent, which can be interpreted as Eq. 1 provides better modeling, owing to its simplicity.

$$d_3 = d_1 - d_2 \quad [5]$$

Experimental data of lipophilicity-based applications and metrics used in medicinal, food, and environmental chemistry.

We also investigated the impact of the ion pair partitioning (P_{IP}^{app}) contribution to lipophilicity-based parameters commonly used in the fields of food, medicinal, and environmental chemistry. Two tests were conducted for food applications. First, we evaluated Eqs. 1 and 2 to reproduce the experimental log$D_{4.5}$ for the partition of bioantioxidants in a oil/water system.[97] Secondly, we collected data on the distribution of spiked drugs in milk products using the pH, pK_a, logP_N and log$D_{6.8}$ reported in the original work.[98,99] However, ion pair partitioning was obtained

from ChemAxon, except for the oxytetracycline (OTET), for which the experimental logP_{IP}^{app} was found in the literature[100] and used to measure the log$D_{6.8}$ using Eq. 1 and Eq. 2.

In addition, as an environmentally relevant application, we obtained experimental pH-dependent distribution data for a series of ionizable compounds on a passive sampler polydimethylsiloxane (PDMS) and water. For this task, monoprotic acids and bases were searched for within the 514 compounds in the article. The experimental pK_a values, log$D_{PDMS/w}$, at several pH values including extreme ranges (from which we were able to obtain logP_N and logP_{IP}^{app}) were provided by the article.[101] Therefore, predictions of the distribution coefficients in the PDMS/water system to pH = 7.4 using Eq. 1 and Eq. 2. were calculated and compared with those reported in experimental work.

Finally, we explored the influence of Eq 1. and Eq. 2 to predict a relevant metric used in medicinal chemistry lead optimization affairs, the lipophilic efficiency (LipE, Eq. 6):[102]

$$LipE = pAct - log\ D_{pH}, \quad [6]$$

where logD_{pH} stands for the distribution coefficient and p*Act* (pIC_{50}, pK_i or pK_b) represents the negative logarithm of biological activity, that is, half maximal inhibitory concentration (IC_{50}, mol/L), inhibitory constant (K_i, mol/L), or binding energy constant (K_b). Here, we searched the literature for ionizable monoacidic or monobasic drug-like molecules with both experimentally determined logD_{pH} measurements and biological activities.[103–114] In some cases, the experimental logP_N, pK_a, and logP_{IP}^{app} were reported, but otherwise they were determined using ChemAxon. The LipE was then simulated using Eq 1. and Eq. 2 and compared to their experimental values.

Machine Learning models to classify the molecules according to the best fit to pH-dependent lipophilicity profiles

Topological and constitutional descriptors were calculated with the software '*rcdk'* package in R[115] while experimental measurements (i.e., logP_N, pK_a, and pH) were added from our dataset. We also added the free energies of hydration and hydrogen bond strengths computed using the new open-source tool '*Jazzy*'.[116] The H-bond donor and acceptor strengths were obtained by calculating the partial charges of the hydrogen atoms and atoms with lone electron pairs, respectively, along with corrective terms. The free energy of hydration was calculated using the

sum of the polar, apolar, and interaction terms. The polar term was derived from the previously calculated H-bond donor and acceptor strengths. The apolar terms consist of the sum of the weighted contributions of the topological surface area, number of rings, and p-orbital counts in the sp and sp^2 atoms. The interaction term consists of a weighted sum of the amount of neighboring H-bond acceptor groups each atom has in a molecule.[116]

We eliminated intercorrelated properties so that no descriptor had a correlation value of $R^2 > 0.6$ (Figure S1 and S2). After this filtration step, two different feature selection methods were tested to choose the best descriptors for the Machine Learning models. First, we performed Welch's *t*-test (**WTT**), which evaluates the statistical difference between the means of two populations that have unequal variances and sample sizes.[117,118] The algorithm calculates the mean of both groups from the binomial conditional for each descriptor. These values were evaluated using Equation 7:

$$t = \frac{\Delta\mu}{\delta_{\Delta\bar{x}}} \qquad [7]$$

where *t* stands for the statistic *t* in Welch's t-test, and Δμ represents the mean difference between data samples from each population (Eq. 1 or Eq. 2 better fits), and the uncertainty value of both groups, which was calculated using the standard deviation of both population samples (Eq.8):

$$\delta_{\Delta\bar{x}} = \sqrt{\left(\frac{s_1}{\sqrt{N_1}}\right)^2 + \left(\frac{s_2}{\sqrt{N_2}}\right)^2} \qquad [8]$$

WTT was performed for each descriptor using R, and the *p*-value was extracted. Features that did not show statistical significance between means ($p > 0.05$) were eliminated. Second, recursive feature elimination (**RFE**) was performed. This iterative feature selection method builds a predictive model using the entire set of descriptors and calculates its importance score (Figure S3). The least important descriptors were removed, and the model was reiterated to achieve maximum performance.[119] This RFE algorithm was programmed using the '*caret'* package in R[120] and tuned via a 5-time repeated *k*-fold cross-validation ($k = 10$). Table 1 shows the descriptors selected using the WTT feature selection method for acids and bases, along with their definitions and target molecules. Table S1 lists the descriptors selected using the RFE method.

Table 1. List of the most influential structural descriptors[95,116,121,122] used for the Machine Learning classification models, their target molecules, and the divergence between the two populations from our dataset were determined using the WTT feature selection method by separating the populations with the conditional $d_3 > 0.2$.

Descriptor	Type	Definition	Target molecules
MDEC.11	Topological CDK descriptor	Molecular distance edge between all primary carbons.	Acids
MDEC.22		Molecular distance edge between all secondary carbons.	Acids
khs.sCH3		Number of -CH_3 fragments in a molecule (Kier and Hall).	Acids
C2SP3		Singly bound carbon atom bound to two other carbons.	Acids
khs.dsCH		Number of =CH- fragments in a molecule (Kier and Hall).	Acids
khs.sNH2		Number of -NH_2 fragments in a molecule (Kier and Hall).	Acids
khs.dssS		Number >S= fragments (sulfones) in a molecule (Kier and Hall).	Acids
HybRatio		Ratio of the number of sp^3-C atoms compared to the sum of sp^3 and sp^2 C atoms.	Acids
C1SP3		Singly bound carbon atom bound to one other carbon.	Acids
nRings7		Number of 7-membered rings	Bases
khs.aaNH		Number of Ar-NH-Ar fragments in a molecule (Kier and Hall).	Bases
ATSc3		Autocorrelation topological distance weighed by charge calculated at every 3-atom distanced segment. Moreau-Broto autocorrelation descriptor 3 using polarizability	Bases
Alogp2	Constitutional CDK descriptor	$(\log P)^2$ value calculated with a QSAR method (Ghose & Grippen $\log K_{o/w}$).	Acids & Bases
delta	Experimental descriptor	δ (acids) = pH - pK_a δ (bases) = pK_a - pH	Acids & Bases
CH_strength	*Jazzy* calculation	C-H donor strength predicted with the *Jazzy* calculations.	Acids

Logistic Regression Classification

A logistic regression (LR) is a simple classification statistical model that provides a binary response to the distribution of the input data among a specific descriptor. The simplest regressions fit the distributions of data to a sigmoidal function, where the input values are given a probability value, which is then classified into one of the two classes based on a cut-off value. We firstly performed a feature selection process specific for logistic regressions by using the '*bestglm'* package in R[123] which evaluates through *n* iterations, which combination of descriptors gives the best fitted regression through the *leaps* algorithm.[124] This package evaluates the weight of each descriptor by linearizing the sigmoidal function and giving a slope value and standard error for each parameter like a multiple linear regression model (Equation 9).

$$\ln\left(\frac{f(x)}{1-f(x)}\right) = \sum_{i=1}^{n} c_i x_i + \mathrm{b} \qquad [9]$$

The '*bestglm'* package drops the parameters, where $c_i \to 0$. The algorithm iterates the sigmoidal fit using Equation 8 *n* times until it finds the combination of descriptors in which the parameters have the smallest standard error.[123] This feature selection process was performed separately for acids and bases because the descriptors have different behaviors for each type of molecule.

Figure 3 shows a flowchart of the modelling process where the dataset was divided into acids (113 entries) and bases (100 entries). Here, acids and bases were modeled separately and labeled as **Models A** and **B**, respectively. Zwitterions (12 entries) were not considered for the Machine Learning predictions because of their small sample size and because further lipophilicity modeling will be performed for these molecules (see Results and Discussion section). Multiple logistic regressions were performed for the training sets based on previously collected descriptors. Predictive models were programmed using the '*caret'* package. (see Figure 3). The test sets were evaluated using both models. The performance of Models A and B was evaluated using confusion matrices (see Table S2), which are widely used to evaluate classification models.[125] The confusion matrices tabulate the number of true positives (**TP**), false positives (**FP**), true negatives (**TN**), and false negative (**FN**) predictions, along with the sensitivity, specificity, and accuracy of the models. Sensitivity determines the ability of the model to detect events of the positive class; that is, it indicates the predictive performance of the molecules of the log$D_{\mathrm{Eq.2}}$ population (Equation 10). On the other hand, the specificity indicates the performance of the model in detecting the negative

class, which in this case are the molecules of the log$D_{\text{Eq.1}}$ population (Equation 11). The accuracy indicates the overall performance in detecting false positives and false negatives (Equation 12).

$$Sensitivity = \frac{TP}{TP + FN} \quad [10]$$

$$Specificity = \frac{TN}{FP + TN} \quad [11]$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad [12]$$

Models A and B were tested further using an external set. The experimental lipophilicity measurements made by Disdier *et al*.[45] consisted of 69 data entries of small molecules with 38 acids, 16 bases, and 15 zwitterions, the latter being discarded for our analysis. To further check the robustness of our models, a second external set of amino acid analogs was evaluated[126], consisting of 8 entries of histidine (basic amino acid) and 10 entries of tyrosine (acidic amino acid). Then, we evaluated the performance of Models A and B for this dataset using confusion matrices (see Table S3-S4).

10.1002/cphc.202300548

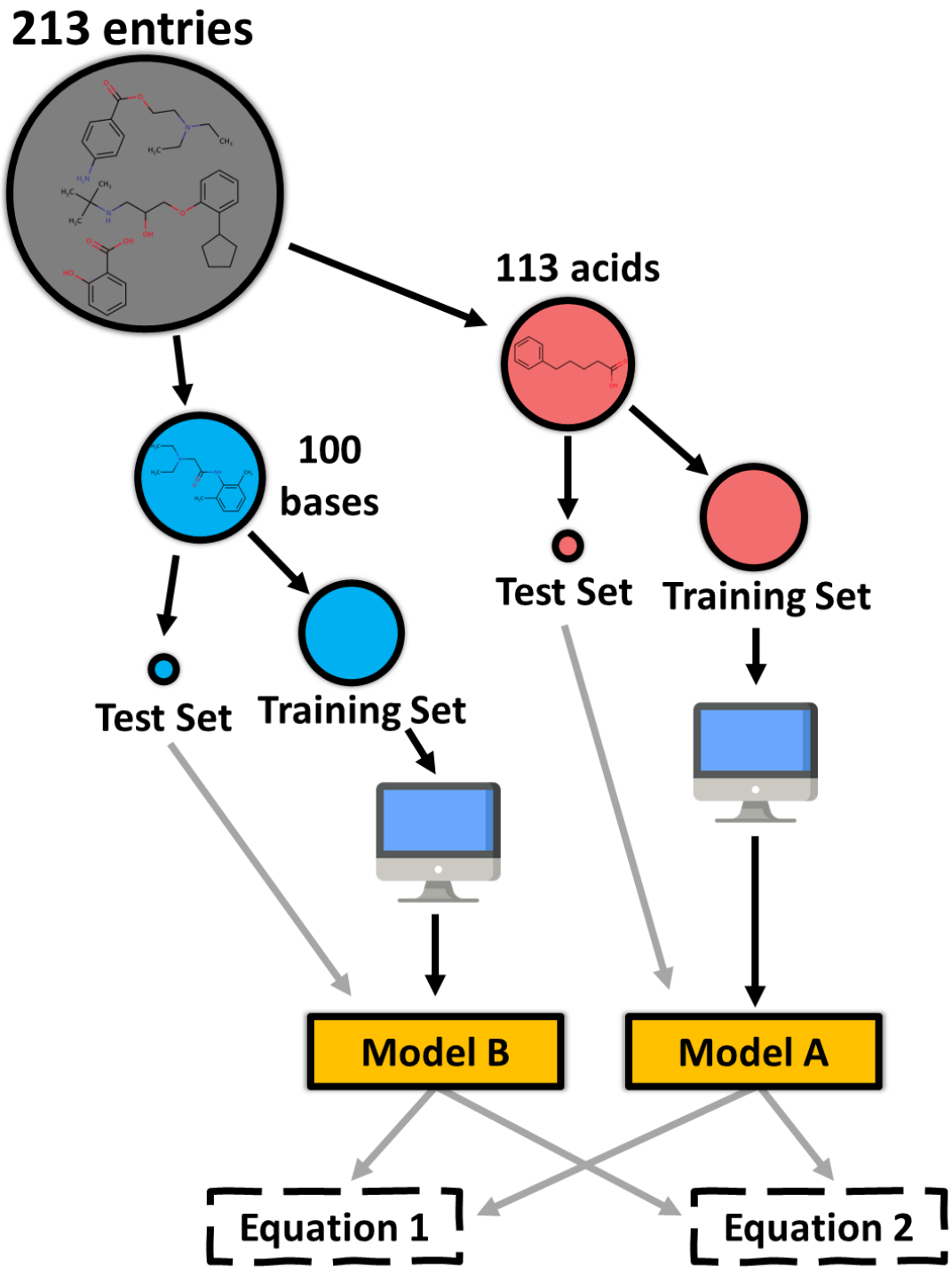


Figure 3. Graphical representation of the data classification and sampling of our dataset to create our predictive multiple logistic regression model using topological, constitutional, and experimental descriptors. The computer icon was made by Freepik from www.flaticon.com.

Random Forest Classification

Decision trees are a simple visual method for evaluating or classifying data, where each node consists of a variable in the dataset. Each node leads to a leaf in which the desired output is issued. A random forest is a combination of decision trees, which are randomly sampled, and the nodes are randomly organized.[127] We split our dataset into training-, and test sets, as shown in Figure 2. In this case, Models A and B consisted of random forest classifications (**RFC**) performed with the '*randomForest*' package in R.[128] Both models were previously refined using the *tuneRF* function within the package, which chooses the optimal *mtry* variable. This value indicates the number of features selected at each split in each decision tree, where *mtry* = 2 provided the best prediction for both models (number of trees = 500, see Supporting Information Figure S4). The importance of each descriptor in both models was evaluated through the mean decrease in the Gini impurity index using the *MeanDecreaseGini* function (Figure S4).

The best lipophilicity profile fit for the acidic and basic tests and external sets was predicted with Models A and B, respectively. The performance of each prediction was evaluated using confusion matrices (see Tables S5-S7) and their respective sensitivity, specificity, and accuracy calculations (Eqs. 10-12).

Support Vector Machine Classification

A Support Vector Machine (**SVM**) algorithm works by dividing training data into two categories, either by linear or nonlinear classification; new data are then assigned to one of the two classes. The model separates the data by finding a hyperplane that maximizes the gap between categories. In the case of linear classification, the space is two-dimensional, making the hyperplane a linear function.[129] When the data are not linearly separable, the algorithm performs the kernel trick, which involves increasing the dimensions of the data space. This results in the hyperplane being able to be another function in the original space, such as radial or polynomial, allowing to classify the data in different ways.[130]

We split our datasets in the same manner as the other classification models and set Models A and B as support vector machines given by the *'e1071'* package in R.[131] We decided to compare the performance of using a linear kernel (**SVML**) and a polynomial kernel (**SVMP**). Radial kernels were not evaluated because our binary data do not follow a circular separation in the hyperplane;

therefore, our classifications do not provide an adequate fit. The hyperparameter selection for each model was performed with the *trainControl* and train functions from the *'caret'* package, which executes a *k*-fold cross-validation ($k = 10$ was used), where different values of the parameters were tested and selected, which resulted in the highest accuracy. The best hyperparameters were the function's default parameters: $C = 1$ for SVML and for SVMP, $C = 1$, *degree* = 3, *gamma* =1, and *coef0* = 0. We calculated the accuracy, sensitivity, and specificity of each model using Eq. 9-11, using the results from their respective confusion matrices (see Tables S8-S13). We then compared the confusion matrices of the LR, RFC, SVML, and SVMP models to determine the one that yielded the best results.

Results and Discussion

Performance of pH-dependent lipophilicity profiles in predicting experimental distribution coefficients

One of the main objectives of this study was to assess the extent of the most widely used formalisms in the literature for reproducing experimental pH-dependent distribution coefficients in small molecules. To achive this task, we built a database which consists of experimental pK_a, logP_N, logP_{IP}^{app}, and log$D_{7.4}$ values reported by Avdeef.[29] In addition, we employed experimental entries of 86 molecules from the work of Tsantili-Kakoulidou *et al.,* containing logD_{pH} values at various pH for each molecule as an individual entry.[55] Entries with logD_{pH} values measured in the presence of background salt concentrations greater than 0.15 mol/L were discarded first because of the limited amount of data available.[23,31–40] Secondly, for these few values found, the pK_a was not determined under these conditions, which limited the use of Eq. 1 and Eq. 2. Finally, since it has been reported that both pK_a and lipophilicity can vary by the effect of counterions, this work focused on studies with the highest data availability (physiological salt conditions 0.15 mol/L of NaCl or KCl), leaving higher salt concentrations outside the scope of our study. Unfortunately, the lack of studies that allow the systematic collection of large amounts of data at concentrations higher than 0.15 mol/L, imposes important limitations to this study, mainly in some food and

environmental aspects where high concentrations of salt are used as a preservative[132] and in studies in groundwater contaminated by its high salt content[133], respectively. Consequently, the applications mentioned below were selected under conditions of background salt concentrations of 0.15 mol/L.

Thus, we obtained 225 entries (118 individual molecules) with 113 acids, 100 bases, and 12 zwitterions whose experimental values of pK_a, logP_N, logP_{IP}^{app}, and log$D_{7.4}$ were collected at background salt concentrations of 0.15 mol/L NaCl or KCl. The limited number of data entries in our database lies in the lack of publicly available data of experimental measurements of pH-dependent lipophilicity profiles of molecules, especially the reduced number of apparent ion pair partitioning coefficient measurements in the literature.

The logD_{pH} was calculated using Eq. 1 and Eq. 2 for each molecule at their respective pH values. Figure 4 shows the overall performance of each model by comparing the modeled values with their respective experimental logD_{pH} values. As expected, most of the molecules whose logD_{pH} values were measured under different pH conditions to 7.4, present the largest deviation using Eq. 1 (see Figure 4a, red marks), with highly underestimated predictions. As a consequence, Eq. 1 poorly predicts logD_{pH} values at extreme pH values. On the other hand, the predicted values using the Eq. 2 are significantly better (see Figure 4b), reducing the RMSE by 0.49 log$D$ units, which represents an improvement of 57 % in accuracy.

10.1002/cphc.202300548

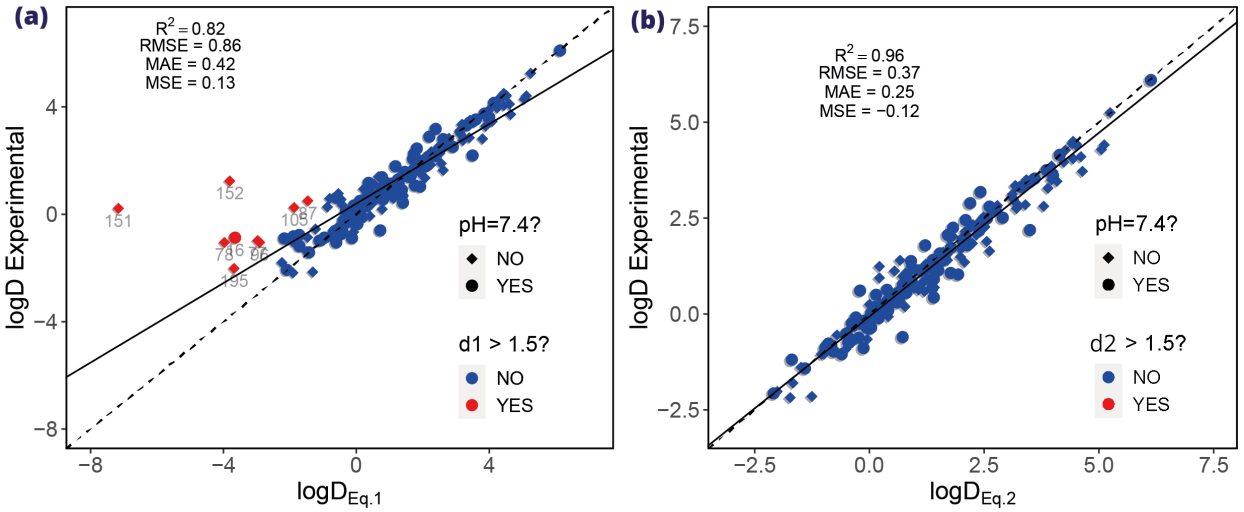


Figure 4. Evaluation of the computed logD_{pH} of our database compared with the experimental values with (a) Eq. 1 and (b) Eq. 2. Rhomboids represent logD_{pH} when the pH is different of 7.4. Red dots and rhomboids highlight compounds with deviations greater than 1.5 log$D$ units. Statistical parameters were calculated using the '*Metrics'* package in R (R^2 = squared Pearson's correlation coefficient, RMSE = root mean squared error, MAE = mean absolute error, and , MSE = mean signed error).

Table 2 shows the reduction of RMSE in log$D$ units of each molecule type using Eq.2 instead of Eq.1. It is observed that our dataset shows a significant improvement (ca. 54 %) in its performance when its distribution coefficient is modeled with log$D_{Eq.2}$ (see Figure S5). Basic molecules showed the greatest improvement, amounting to 66 %, whereas the acid ones 44 %.

Table 2. Decrease in RMSE terms for each type of molecule analyzed within our dataset by comparing the modelled lipophilicities by log$D_{Eq.1}$ and log$D_{Eq.2}$ with their experimental values (see Figure S5).

Type	ΔRMSE[a]	Decrease in RMSE (%)[b]
Acid	0.30	44
Base	0.67	66
All	0.49	57

[a] $\Delta\text{RMSE} = \text{RMSE}\left(\log D_{\text{Eq.1}}\right) - \text{RMSE}\left(\log D_{\text{Eq.2}}\right)$

[b] $\% = \Delta\text{RMSE} \times 100/\text{RMSE}\left(\log \text{D}_{\text{Eq.1}}\right)$

The molecules with the highest deviations in the prediction of experimental logD_{pH} using logD_{Eq1} are displayed in Figure 5. The chemical nature of the outliers is dominated by the presence of ionic species of organic compounds (fractions close to 1) because the distribution coefficients for these compounds were experimentally measured under pH conditions favoring charged species. These deviations correspond to the theoretical frameworks of Eqs. 1 and 2. Thus, the inclusion of the term P_{IP}^{app} in Eq. 2 significantly corrects the prediction. Figure 5 also shows various polyacids with K_a separated by four orders of magnitude, allowing us to analyze the distribution coefficient using the most acidic pK_a. Bases with entries **16, 151-152** have different possible protonation sites, however, these compounds can only be individually protonated at experimentally realistic pH values. On the other hand, acids with entries **77, 78,** and **87** have two deprotonation sites. More complex thermodynamic models can be considered for these molecules,[45] however, as mentioned above, because of the separation of their pK_a, the consideration of P_{IP}^{app} for the carboxylate species with log$D_{Eq.2}$ is enough to remarkably increase the accuracy of the lipophilicity prediction of these compounds to extreme pH where one charged species can predominate over the others.

10.1002/cphc.202300548

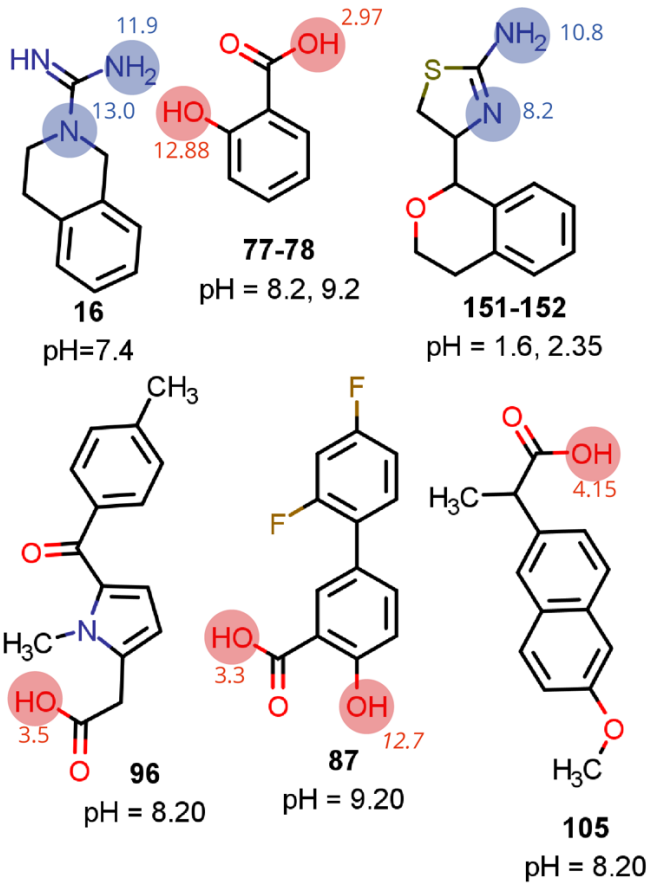


Figure 5. Representation of the molecules with the highest deviations in the prediction of the experimental logD_{pH} using logD_{Eq1}. The pH values of each entry are represented below each molecule. The protonations and deprotonation sites, along with their respective experimental pK_a values, are labeled in blue and red, respectively.[29,55] The pK_a values in italics were calculated using ChemAxon.

We also tested the performance of both formalisms by evaluating the entire pH-lipophilicity profile of individual molecules. To this end, an acidic (Naproxen) and two basic (compound '1774' from Ref 55 and Debrisoquine - compound '16') examples were used. Naproxen and the compound '1774' (see Figure 6a-b) were selected because of the large amount of experimental data available in our database.[56] Therefore, it is better appreciated how the behavior of the experimental lipophilicity profiles fits more closely when evaluated using Eq. 2, particularly at extreme pH values. Additionally, at pH 7.4, the influence of apparent ion pair

partitioning can be observed depending on the chemical nature of the molecule. For instance, Debrisoquine in Figure 6c represents a base with a very high pK_a. Hence, ionic species were more abundant at pH 7.4, which aligned more closely with the lipophilicity profile determined by Eq. 2. These results indicate that to reproduce the pH-dependent lipophilicity profiles of small molecules, it is recommended to use Eq. 2, especially under pH conditions where ionic species of organic compounds are considerably more representative than neutral species.

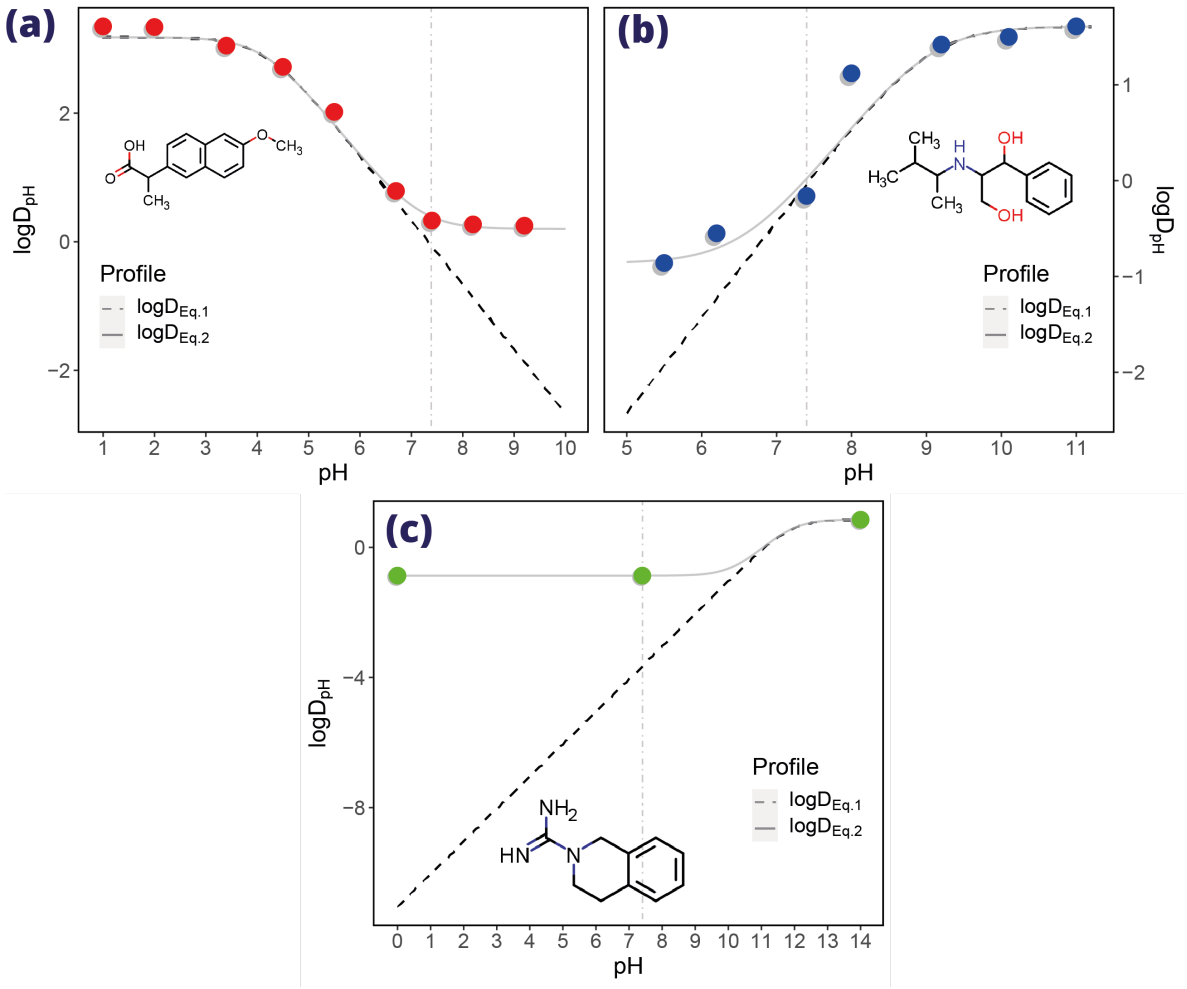


Figure 6. Calculated pH-dependent lipophilic profiles of (a) acidic (Naproxen) and (b-c) basic molecules (compound '1774' from Ref 55 and Debrisoquine) within our dataset. The dashed lines represent the logD_{pH} values calculated using Eq. 1, and the solid line represents the values calculated using Eq. 2. The dots represent experimental logD_{pH} values. A dot-dashed vertical line was placed at pH 7.4.

Use of pH-dependent distribution coefficients in medicinal, food, and environmental chemistry.

Although the distribution coefficient in solvent systems represents only a mimetic for many biological and physicochemical processes, its relevance and successful application in several life sciences fields is undeniable. In this regard, we further investigated the repercussions of the apparent ion pair partitioning of molecules in the applied parameters and metrics where lipophilicity is relevant.

First, owing to the availability of experimental values for pH-dependent distribution coefficients in the olive oil/water system for two bioantioxidants[96], we simulated the log$D_{4.5}$ for these phenolic acids using Eq. 1 and Eq. 2. Table 3 shows that Equation 2 fits best with gallic acid. On the other hand, for caffeic acid, an appreciable error was observed using both formalisms, amounting to almost 1 log$D$ unit. Let us mention, that since both formalisms yield similar predictions, the error does not come from the logP_{IP}^{app}, but in partitions using oil as the organic phase, the penalty for the partitioning of ionic species seems to be higher than that predicted by the difference between pH and pK_a.

Table 3. Experimental and modeled distribution coefficients for the two bioantioxidants to a pH of 4.5 in the olive oil/water system using Eq. 1 and Eq. 2.

Compound	Experimental values (Ref 97)				Calculated log$D_{4.5}$ (Δlog$D$[a])	
	log$P$oil/water	pK_a	logP_{IP}^{app}	log$D_{4.5}$	Eq. 1	Eq. 2
Gallic acid	2.97	4.40	2.34	2.70	2.62 (-0.08)	2.73 (0.03)
Caffeic acid	3.26	4.54	1.70	2.04	2.98 (0.94)	2.99 (0.95)

[a] $\Delta\log D$ = (calc − exp)

Second, previous studies have shown that the distribution of spiked drugs between milk fractions, for example, the curd/whey system, to a pH of 6.8, can be properly mimicked through the *n*-octanol/water distribution coefficient (log$D_{6.8}$) using Eq. 1 (see Figure 7a).[98,99] Despite the excellent results obtained using Eq. 1, we were interested in investigating whether the use of Eq. 2 further improves the observed model (see Figure 7b).

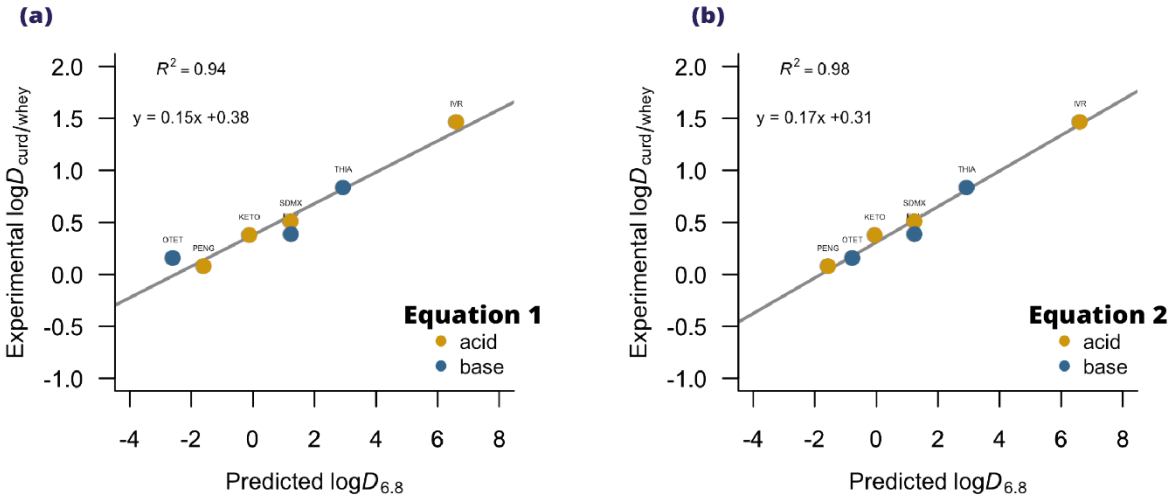


Figure 7. Comparison between *n*-octanol/water log$D_{6.8}$ using Eq. 1 (a) and Eq. 2 (b) and the experimental distribution of spiked drugs between the curd/whey milk fractions. Drugs that represent each acronym in the plot are listed in Table 4.

Table 4 reports the data used in the previous report[98,99] as well as the logP_{IP}^{app} of the tested molecules picked up from experimental measurements from the literature when available; otherwise, this parameter was simulated using predictive software such as ChemAxon.[94,100] The predicted *n*-octanol/water distribution coefficients for spiked drugs to pH = 6.8 using Eq. 2 (see Fig 7, right) improved the correlation from 0.94 to 0.98 and showed an improved linear regression between both descriptors (see the linear equations in Fig. 7). Let us mention that previous studies have shown that experimental errors in the experimental determination of partition coefficients can be as low as 0.3 logP units[30] so although this is a small improvement, the one presented in

Figure 7 may be representative in terms of the involvement of ionic species. In more detail, the improvement resides precisely in two compounds that had an experimental value of $\log P_{IP}^{app}$ (ketoprofen and oxytetracycline). This observation highlights the importance of experimental measurements of ion pair partitions but also calls for more experimental work focused on these issues, taking into account the scarce values available in the literature and the difficulty of finding them in public databases.

Table 4. Experimental distribution of spiked drugs between curd/whey milk fractions and predicted *n*-octanol/water distribution coefficients for spiked drugs at pH = 6.8 using Eq. 1 and Eq. 2.

Compound	Original data from Refs. 98 and 99			$\log P_{IP}^{app}$	Calculated $\log D_{6.8}$	
	$\log P_N$	pK_a	$\log D$ curd/whey (pH = 6.8)		Eq. 1	Eq. 2
Penicillin G (PENG)	1.67	3.53	0.08	-2.75	-1.60	-1.57
Sulfadimethoxine (SDMX)	1.48	6.91	0.51	0.12	1.23	1.24
Ketoprofen (KETO)	2.81	3.88	0.38	-0.95[a]	-0.11	-0.05
Ivermectin B1a (IVR)	6.61	12.47	1.47	1.55	6.61	6.61
Oxytetracycline (OTET)	-1.60	7.75	0.16	-0.74[b]	-2.60	-0.78
Erythromycin A (ERY)	2.83	8.38	0.39	-0.89[b]	1.24	1.24
Thiabendazole (THIA)	2.93	4.08	0.84	1.28	2.93	2.93

[a]Experimental data reported in our database in Ref 56. [b] Experimental data reported in Ref. 99.

10.1002/cphc.202300548

Additionally, in environmental chemistry research, passive equilibrium sampling of dissolved contaminants in water has been studied using polymer polydimethylsiloxane (PDMS) as an absorbent phase. This hydrophobic passive sampler can extract ionizable compounds from sediments and suspended particulate matter in a pH-dependent manner.[101] The authors tested the partitioning of ionizable compounds between PDMS and water ($\log D_{PDMS/w}$) at different pH values. Thus, $\log D_{PDMS/w}$ measurements at extreme pH were considered as $\log P_N$ or $\log P_{IP}^{app}$ depending on the acidic or basic nature of each molecule. These values were used to calculate $\log D_{Eq.1}$ and $\text{Log}D_{Eq.2}$ in order to reproduce the experimental $\log D_{PDMS/water}$ measurements to a pH of 7.4 (see *environmental.csv* on the github repository).

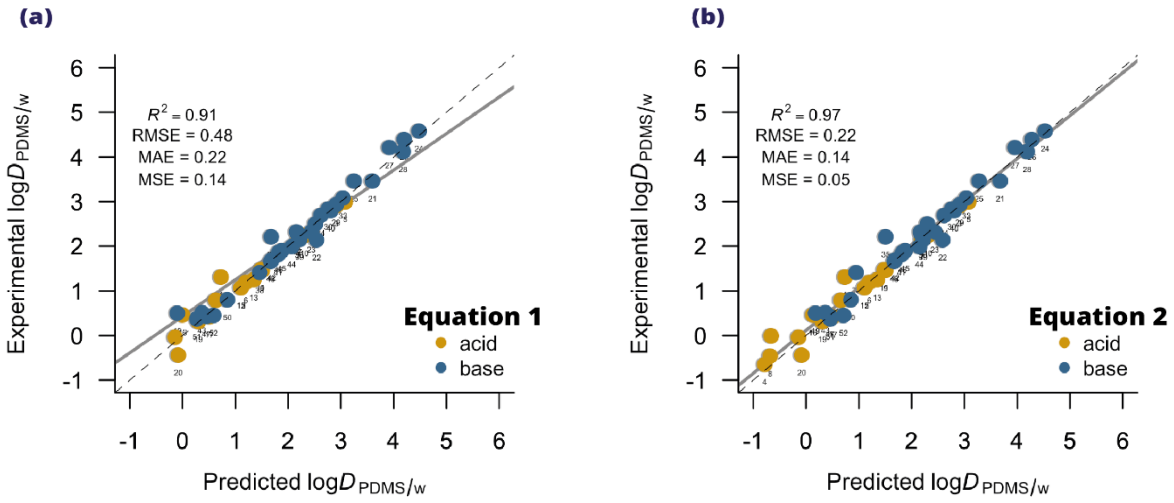


Figure 8. Comparison between PDMS and water $\log D_{7.4}$ using Eq. 1 (left), and Eq. 2 (right), and the experimental $\log D_{PDMS/w}$ for a series of 52 ionizable compounds.

Figure 8 shows that the consideration of the apparent ion pair partitioning significantly improved the correlation between the experimental and predicted values, where the RMSE decreased in 0.26 log*D* units, representing an appreciable improvement of 54 %. This application demonstrated that the thermodynamic equilibrium derived in Eq. 2 applies to partitions other than the *n*-octanol/water system, such as in the PDMS/water phases. The presence of some free silanol groups, combined with the highly hydrophobic polymeric chain might create a suitable environment for ionized-organic species in the PDMS phase through a combination of hydrogen bonds from the terminal -OH groups and dispersion non-covalent forces from the polymeric chains.[101,134]

Finally, an important metric that has been increasingly applied in drug discovery and medicinal chemistry lead optimization endeavors is the lipophilic efficiency (LipE, see Eq. 6). LipE relates the binding affinity and lipophilicity of a compound, which creates a significant parameter for estimating druglikeness.[102] A proper interval of lipophilicity at physiological pH (log$D_{7.4}$), usually between 1 and 3, underpins the desired ADME properties and dose; therefore, improving potency without excessively increasing lipophilicity is of vital importance in drug discovery optimization programs. Table 5 compiles the LipE values obtained in the literature using strictly experimental *pAct* (pIC_{50}, pK_i or pK_b) and lipophilicity at physiological pH, the latter due to the impossibility of finding data at other pH values in the literature.

Table 5. Experimental lipophilic efficiency (LipE) for compounds reported in the literature and predicted lipophilic efficiency at pH = 7.4 using Eq. 1 and Eq. 2. The number representing the compounds in each original work is placed in column 'Id'.

Compound	Id (Ref)	Exp. *pAct*	Exp. log$D_{7.4}$	Exp. LipE	logP_N	pK_a	logP_{IP}^{app}	LipE Eq. 1	LipE Eq. 2
1	6 (112)	8.5	1.4	7.1	2.07	8.42	-0.20	6.51	6.51
2	8 (112)	7.9	1.3	6.6	1.88	8.42	-0.39	6.01	6.01
3	9 (112)	8.1	1.7	6.4	2.37	8.42	0.09	5.73	5.72
4	10 (112)	7.8	2.1	5.8	2.29	8.42	0.03	5.60	5.60
5	11 (112)	7.9	2.3	5.6	3.59	8.42	1.32	4.34	4.34
6	12 (112)	7.8	3.5	4.3	3.43	8.42	1.16	4.43	4.43
7	Rosuvastatin (109)	9.0	-0.3	9.3	1.90	4.27	-1.63	10.23	10.08
8	Pravastatin (109)	7.1	-0.2	7.4	2.18	4.20	-2.41	8.15	8.13
9	Fluvastatin (109)	6.6	1.9	4.7	4.17	4.30	0.18	4.65	4.65
10	8 (108)	8.4	3.0	5.4	4.50	3.84	1.67	7.46	6.66
11	9 (108)	7.7	2.4	5.3	4.99	3.89	1.42	6.20	5.93
12	12 (108)	8.2	2.5	5.7	5.01	4.07	1.41	6.47	6.28
13	13 (108)	8.3	3.2	5.1	5.61	3.92	2.67	6.17	5.52
14	14 (108)	8.3	2.0	6.3	4.23	4.08	2.43	7.39	5.86
15	8 (107)	6.4	2.7	3.7	2.77	7.90	1.50	3.71	3.7
16	10 (107)	5.9	2.3	3.6	2	7.00	0.75	4.43	4.37
17	11 (107)	6.9	2.3	4.7	2.3	7.60	1.00	4.87	4.86
18	19 (107)	4.8	0.1	4.8	0.43	9.50	-2.00	4.43	4.43
19	Indinavir (104)	9.1	2.9	6.2	2.92	6.20	-2.42	5.15	5.15
20	3 (104)	8.0	3.2	4.9	4.5	8.97	0.77	5.18	5.17
21	Crizotinib (104)	8.1	1.9	6.1	4.28	9.40	-0.38	5.82	5.81
22	8l (112)	6.4	1.7	4.7	1.78	5.66	-1.96	6.42	6.42
23	22 (102)	7.0	3.6	6.4	4.24	8.87	0.43	5.28	5.28
24	7a (113)	9.2	2.8	6.3	2.65	9.60	-1.13	7.66	7.65
25	7j (113)	5.6	1.0	5.5	3.38	9.70	-0.41	7.03	7.01
26	7k (113)	7.9	1.6	6.1	3.04	9.70	-0.75	7.34	7.32
27	7m (113)	6.8	2.1	6.6	2.22	9.70	-1.57	8.18	8.16
28	7s (113)	7.9	2.4	6.0	2.89	9.70	-0.90	7.53	7.52

Figure 9 shows the LipE simulated using Eq 1. and 2 compared to their experimental LipE values. Eq. 2 again shows favorable statistical parameters compared with Eq.1, improving the RSME in log *D* units by 11 %. Let us mention that the reproduction of LipE in both cases was not very satisfactory, this may be mainly due to the lack of experimental data of logP_N values for these compounds, but in particular to the use of simulated logP_{IP}^{app} values. Although, tools such as ChemAxon have presented very good results in partition coefficient predictions, the molecules included here belong to novel compounds reported in medicinal chemistry articles with new chemical spaces that may impact the performance of predictive tools, especially in logP_{IP}^{app} which to the best of our knowledge, this is the first work in reporting a free database for this parameter taken from several reports in the literature.

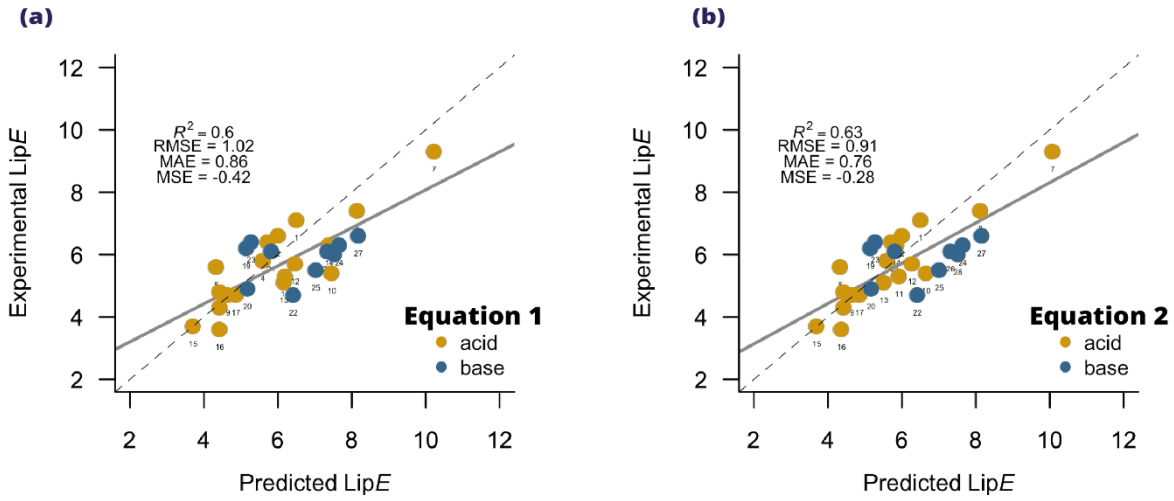


Figure 9. Comparison of the lipophilic efficiency (LipE) using Eq. 1 (left), and Eq. 2 (right), and the experimental lipophilic efficiency for a series of 28 ionizable compounds.

To summarize, it can be noted that the inclusion of apparent ion pair partitioning can improve metrics in the modeling of various descriptors where lipophilicity is crucial to simulate biological or artificial environments of higher complexity. Of special interest, we demonstrated that these formalisms can be applied to systems beyond the classical *n*-octanol/water system. The improvements studied ranged from 4 to 54 % in terms of RMSE (log*D* units) and depended mainly on the pH at which the system was being simulated, pK_a, and lipophilicity of the molecules.

Machine Learning models to guide the choice of pH-dependent lipophilicity profiles as a function of molecular properties.

The application of Eq. 1 offers significant advantages due to its simplicity of implementation. However, the preceding sections emphasize the importance of employing Eq. 2 in various scenarios. It is important to note that the application of this formalism is constrained by the availability of experimental data for partitioning of ionic species and simulations that ensure adequate accuracy. This led us to propose a model that will help as a guide to discern cases in which the simplified model represented in Eq. 1 can be used, or the consideration of the apparent ion pair partitioning is mandatory, as in Eq.2.

Consequently, one of the aims of this study was to develop a classification algorithm that can differentiate whether the lipophilicity profile of a molecule can be better predicted with log$D_{Eq.1}$ or log$D_{Eq.2}$. However, a significant number of entries indicate that both formalisms compute a similar result compared to their experimental values by yielding d_3 values close to 0 (see Figure S6a). Let us note that we focus on the specific cases with a significant improvement when the P_{IP}^{app} of molecules is considered. Therefore, we decided to delimit the conditional d_3, indicating that if a molecule exceeds a certain value of d_3, it is important to consider its apparent ion pair partitioning to predict its lipophilicity. We tested d_3 values between 0.1-1 and picked the optimal value based on two main parameters. First, considering that our set was small because we used strictly experimental values in our database, we seek that the population of molecules that

best fit with log$D_{Eq.2}$ should be at least 10 %. Then, there should be a sufficient number of descriptors that have statistically proven divergence by WTT ($p < 0.05$). Thus, Machine Learning algorithms have a larger number of parameters to create predictive models with higher accuracy. Consequently, the delimiter '0.2' showed an adequate balance between these two parameters and was selected as our cut-off value (see Figure S6b). Molecules with values of $d_3 > 0.2$ showed an improvement in lipophilicity modeling using Eq. 2. On the other hand, entries that had negative d_3 values or that fell into the range $0.2 < d_3 < 0$ were classified as molecules where the difference between both models was negligible, and thus were classified as better fitted using log$D_{Eq.1}$ due to its easy implementation (it does not depend on P_{IP}^{app} , resulting in less computational effort and fewer experimental parameters). Higher thresholds significantly decreased the population in log$D_{Eq.2}$, while lower values reduced the structural divergence between molecules in log$D_{Eq.1}$ and log$D_{Eq.2}$, making it more difficult to find descriptors that can differentiate between both populations. The value '0.5' was also tested because a local maximum of descriptors with $p < 0.05$ was observed at this point (see Figure S6b). Furthermore, this value is of experimental interest, because logP_N measurements of substances with different techniques tend to vary by less than 0.5 log$P$ units (using the Shake-Flask method as a reference), which is considered as a parameter indicating that the experimental techniques are not equivalent.[30] However, this extreme value and the descriptors selected (see Table S14) showed poor performance in the ML models tested, especially with External Set 1 (see Figure S7). This phenomenon can be explained because this d_3 delimiter has a very small log$D_{Eq.2}$ population, thus the datasets are extremely unbalanced and the robustness of the models is reduced. On the other hand, the accuracy of experimental methods, even using different techniques, rounds at values less than 0.2 log$P$ units.[30] Therefore, we continued to train the ML models using the $d_3 > 0.2$ cut-off value to determine tendencies among the selected descriptors via the feature selection methods and to evaluate the performance of the ML algorithms.

Figure 10 shows the distribution of the molecules in our database, classified using the criterion $d_3 > 0.2$ as a binary descriptor. Most entries can be computed using log$D_{Eq.1}$ with satisfactory results. However, we observed that 25 acids and 10 bases showed a clear improvement within our d_3 threshold by modeling lipophilicity with log$D_{Eq.2}$.

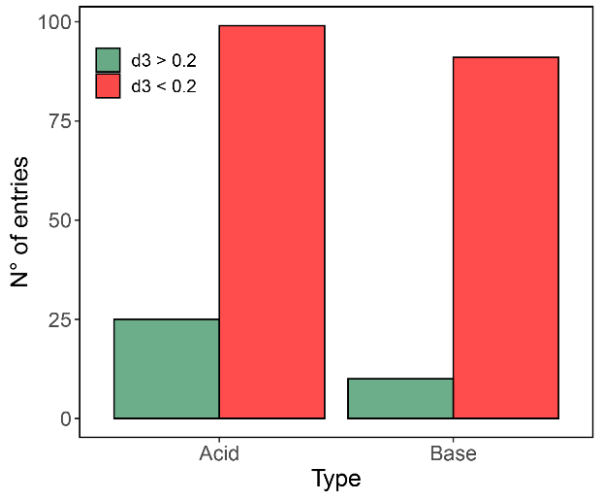


Figure 10. Distribution of acid and basic entries from our dataset as a function of their d_3 values.

We obtained several structural and physicochemical descriptors of the molecules to identify considerable divergence between populations. First, our database was split into acids and bases, and then into the training and test sets. The '*rcdk'* package in R was used to look through the descriptors, along with the *Jazzy* calculations of energies of hydration and hydrogen-bond strengths and the experimental descriptors. The selected feature selection methods showed a wide range of diverse descriptors (see Tables 1 and S1). We performed a Welch's *t*-test on our descriptors (WTT), which analyzes the divergence between populations relative to the variances of the two groups.[117] This test was selected over a Student's *t*-test because of the divergence of sample sizes (Figure 10) and variances between groups (Figure 11-12).[118] The WTT descriptors provided acceptable accuracies (see Figure S8).

An iterative feature selection method was also tested using the RFE model. The algorithm achieved better performance when the 14 most important variables for acids and the nine most important variables for bases were maintained. The importance of each descriptor posed by the RFE is shown in Figure S3. Good results were obtained when these descriptors were implemented during the training of the Machine Learning models. However, the accuracy decreased significantly when the Test and External Set 1 were evaluated (see Figure S8c-d), indicating that these descriptors did not generate a sufficiently robust model, or that the large number of chosen descriptors (see Table S1) may overfit the data. Therefore, we selected WTT descriptors to analyze

the tendencies of the molecules in each population and to evaluate the overall performance of the Machine Learning algorithms that we developed.

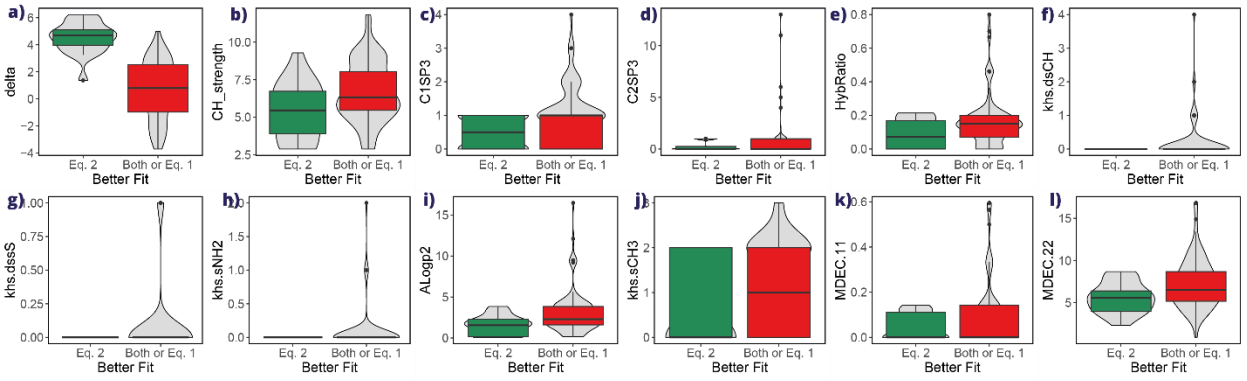


Figure 11. Violin plots of the distribution of the acidic molecules in our dataset along the selected descriptors for the acids ((a) *delta*, (b) *CH_strength*, (c) *C1SP3*, (d) *C2SP3*, (e) *HybRatio*, (f) *khs.dsCH*, (g) *khs.dssS*, (h) *khs.sNH2*, (i) *Alogp2*, (j) *khs.sCH3*, (k) *MDEC.11*, and (l) *MDEC.22*). Distributions are separated between acids and bases and classified by the binary operator $d_3 > 0.2$ (green) and $d_3 < 0.2$ (red).

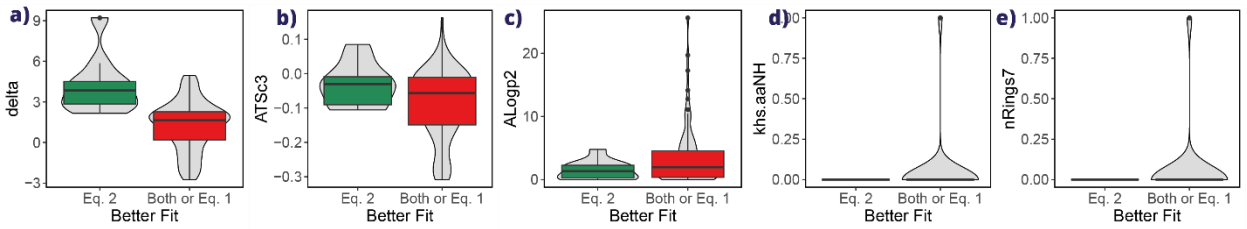


Figure 12. Violin plots of the distribution of the basic molecules in our dataset along the selected descriptors for the bases (a) *delta*, (b) *ATSc3*, (c) *Alogp2*, (d) *khs.aaNH*, and (e) *nRings7*). Distributions are separated between acids and bases and classified by the binary operator $d_3 > 0.2$ (green) and $d_3 < 0.2$ (red).

Figure 11 and 12 show the selected descriptors for acids and bases, respectively, used to train our classification ML models. These descriptors showed statistically significant divergence between the means of both populations among the 180 descriptors tested for acids and bases. Both acidic and basic compounds showed significant differences in their means ($p < 0.05$ in WTT test) for the *delta* and *Alogp2* descriptors (Table 1). The descriptor *delta* was calculated at the respective pH of each entry for acids and bases. As expected, this descriptor proved to be the most important in every test carried out in this regard (see Figures S3-S4), as it correlates with the prominence of ionic species in both phases. Therefore, the apparent ion pair partitioning becomes more significant for entries with higher *delta* values (Figures 11a and 12a). This result is very promising, because despite being an experimental descriptor, there are computational methods to determine pK_a that include first-principles models[135–138] as well as machine learning tools[139,140].Thus, the descriptor *delta* can be automated and easily used to classify molecules according to the lipophilicity formalisms analyzed here. In fact, the root-mean-square error (RMSE) between predicted pK_a values using the software ChemAxon and experimental data in our database is just 0.58 log units and the squared coefficient of determination (R^2) of 0.95 (see Figure S9)

The *ALogp2* descriptor consists of a 3D-QSAR model by Ghose & Crippen (1986), which predicts a square value of the clogP_N value by analyzing the presence of 110 structural fragments within the molecules.[95] Figure 11i and 12c show that molecules with hydrophobicity close to log$P$ = 0 (with lower *Alogp2* values) tend to fit best with log$D_{Eq.2}$. Although water and *n*-octanol are not miscible, a small amount of water can dissolve in octanol at room temperature (~ 2.9 mol/kg).[141] These hydrophilic molecules might be dragged by the dissolved water to the octanol phase along with the ionic species; thus, the apparent ion pair partitioning would have a higher importance in these molecules.

This affinity for water, at least for acidic compounds, was further demonstrated using the *CH_strength* descriptor (Figure 11b). This descriptor, calculated by *Jazzy*, predicts the hydrogen-bond donor strength in carbon atoms.[116] The smaller *CH_strength* values indicate that for entries with $d_3 < 0.2$, H-bond donors are not primarily found on carbons. Instead, they are found on other more electronegative heteroatoms. Thus, by weakening the X-H covalent bonds through H-bonds, the possibility of ionization of these species in both water and *n*-octanol increases. Figure 11e,k-l presents other important descriptors for acidic compounds such as *MDEC.11*, *MDEC.22*, and *HybRatio.* The *MDEC.11* and *MDEC.22* descriptor consists of a relationship between the number

of primary (*MDEC.11*) and secondary (*MDEC.22*) carbons in the molecule and the squared average atomic distance between these atoms.[121] Whereas, *HybRatio* is the number of sp^3-C atoms compared to the sum of sp^3 and sp^2 C atoms. Eq. 2 works better for acidic substances with low values of these descriptors, which considers together the values of *Alogp2*, allows us to infer that small and rigid ionizable molecules with instaurations or aromatic systems need considering the P_{IP}^{app} to obtain an accurate prediction of $\log D_{pH.}$

Similarly, for basic compounds, higher values of the *ATSc3* descriptor are associated with the consideration of the P_{IP}^{app} for modeling pH-dependent lipophilic profiles on basic molecules (Figure 12b). This descriptor is related to the high molecular polarizability, which agrees with the pattern of small molecules in the presence of polar atoms such as nitrogen. Therefore, the apparent ion pair partitioning effect should be considered for these small, rigid, and unsaturated molecules, which present a significant proportion of ionic species in the aqueous phase. It has been previously shown that the P_{IP}^{app} of molecules may mechanistically occur via a simple ion-transfer reaction.[142] Thus, it is more plausible that small and compact molecules have a more prominent P_{IP}^{app} because of the lower energetic cost of transferring to the cavity of the ion they replace.

After establishing a distinct division between the two populations and applying an appropriate feature selection method, Models A and B (see Figure 3) were trained using the logistic regression (LR), random forest classification (RFC), and support vector machine (SVM) algorithms. A training set for acidic and basic molecules was used for each model and was evaluated using a test set consisting of 20% of the population. In addition, the two external sets were validated using the experimental data obtained by Disdier et al. (External Set 1)[45] and Fauchère and Pliška (External Set 2).[126] Predictions were made to determine which formalism best modeled the lipophilicity of the inputs, and the results were collected in confusion matrices. The performance of each marker was evaluated by calculating its accuracy, specificity, and sensitivity. Figure 13 shows the results of the calculations of the four algorithms for the test and external sets of acidic and basic molecules.

10.1002/cphc.202300548

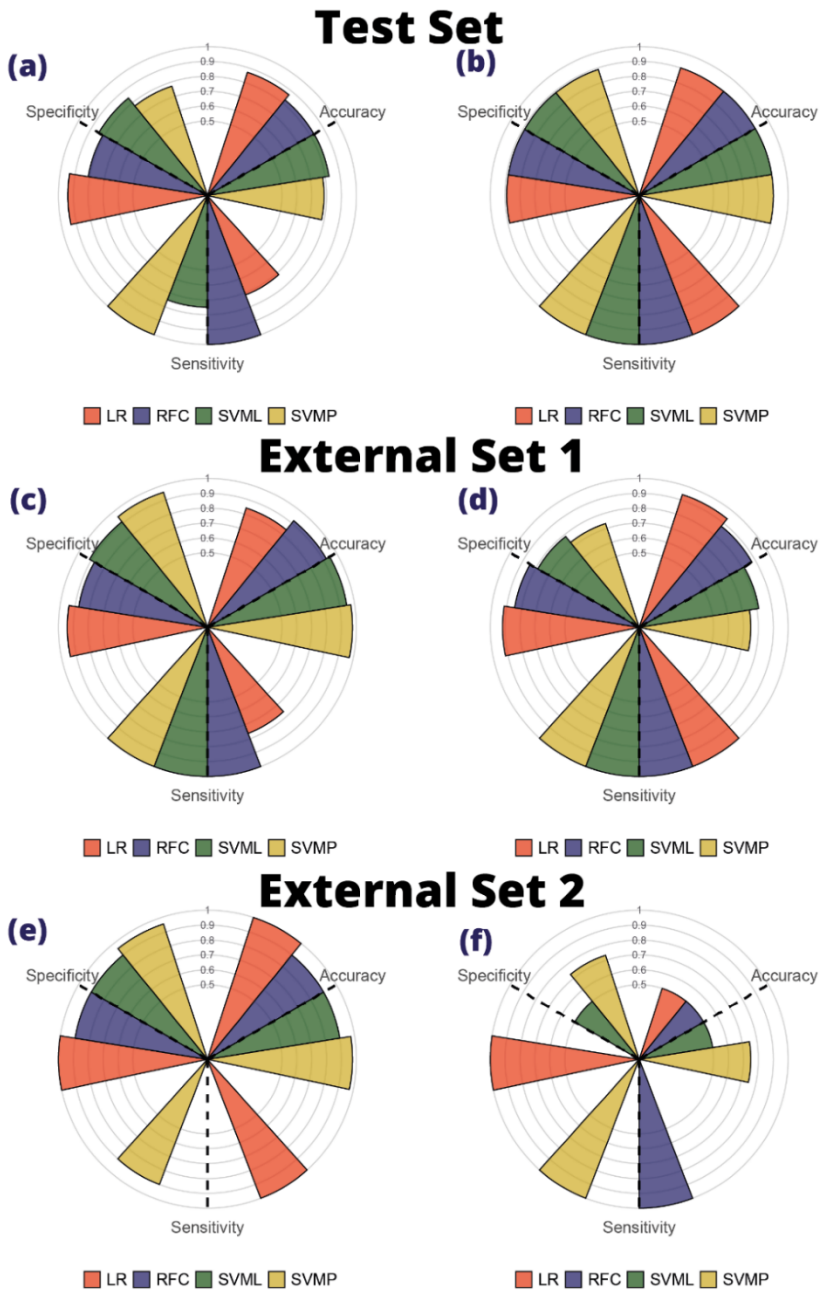


Figure 13. Accuracy, sensitivity, and specificity of each ML model evaluated in this study for acidic (a,c,e) and basic (b,d,f) entries within the test and external sets by defining our populations with the conditional $d_3 > 0.2$. Descriptors were selected using the WTT method. Accuracies, sensitivities, and specificities were calculated with Eqs. 10-12 based on the results of each confusion matrix (Tables S2-S13)

10.1002/cphc.202300548

It is observed that most of the calculated accuracies for our test set have high values (between 0.8 and 0.95), denoting that these classification models manage to distinguish relatively well which molecules best fit with log$D_{Eq.1}$ and log$D_{Eq.2}$. However, the sensitivity decreased in the test set of acidic molecules, indicating that the models had difficulties in detecting molecules that fit log$D_{Eq.2}$ (Figure 13a). External Set 1 exhibited good performance, with all models showing similar accuracies, sensitivities, and specificities to those evaluated in the Test Set (Figure 13c-d). Additionally, External Set 1 mainly comprises more hydrophobic molecules than our dataset, as most molecules have logD_{pH} values < 0 (Figure S10). This demonstrates that our models exhibit high robustness, even when dealing with species belonging to slightly different chemical spaces. External Set 2, associated with capped amino acids as reported by Fauchère and Pliška[126], obtained divergent results. On the one hand, the pH-dependent values of *N*-Acetyl-*L*-tyrosine amide were predicted with excellent metrics, especially using the LR and SVMP models, because our training set had a representative number of molecules with phenolic groups. On the other hand, in the case of *N*-acetyl-*L*-histidine amide, the results were very poor due, at least in part, to the fact that our set had few bases in relation to the acids that best fit Eq. 2, mainly because there was no imidazole fragments present in our set of bases, thus limiting the performance of our models.

Conclusions

Lipophilicity is undoubtedly the most widely used and important descriptor in the early stages of drug discovery and development. Additionally, it is a crucial descriptor in substance risk assessment and in areas such as adsorption in materials, catalysts, food chemistry, and computational biology. There are multiple tools to determine this descriptor, mainly for neutral molecules (logP_N). For substances with ionizable groups, two formalisms are commonly used to determine the distribution coefficient (logD_{pH}), being the simplest pH correction model is the most widely used. However, previous studies carried out on specific and small molecule sets recommend considering the effect of the apparent ionic compounds (P_{IP}^{app}) because it has a negative impact on the accuracy of computing lipophilic profiles when charged species or related species are ignored. Our study, which was based on a larger amount of data and strictly on experimental values, validated the observations presented in previous studies. We have also evidenced the impact of P_{IP}^{app} on the prediction of both the experimental lipophilicity profiles of small molecules and experimental lipophilicity-based applications and metrics such as lipophilic efficiency (LipE), distribution of spiked drugs in milk products, and pH-dependent partition of water contaminants in synthetic passive samples such as silicones. Our findings show that better predictions are obtained by considering the apparent ion pair partitioning, whereas ignoring its contribution can lead to inadequate experimental simplifications and/or computational predictions.

Finally, we developed machine learning algorithms using logistic regression, random forest classification, and support vector machine models to determine from molecular structures in which cases the P_{IP}^{app} should be considered. The results indicate that small, rigid, and unsaturated molecules with logP_N close to zero, which represent a significant proportion of ionic species in the aqueous phase, are better modeled using the formalism that takes into account the apparent ionic compounds (P_{IP}^{app}).

Although we are aware of the molecular complexity of the species that can be included in the computational determination of the apparent ion pair partitioning (P_{IP}^{app}), parameterization or training of models using experimental values of P_{IP}^{app} can help alleviate the restricted application of formalisms that include this effect. Finally, our findings can serve as guidance to the scientific community working in early-stage drug design, food, and environmental chemistry who deal with

ionizable molecules, to determine a priori which lipophilicity profile should be used depending on the structure of a substance in research efforts. Future studies will address the influence played by the apparent ion pair partitioning (P_{IP}^{app}) on the pH-dependent lipophilic profiles in more complex systems such as zwitterionic and peptides.

Data availability

All codes are hosted at: https://github.com/cbio3lab/Lip_profiles. The database constructed in this work can be consulted in the reference 56.

Author contributions

William J. Zamora: conceptualization, methodology, validation, data curation, writing – original draft, writing – review & editing. Esteban Bertsch: methodology, validation, writing – review & editing. Sebastián Suñer: methodology, validation, writing – review & editing. Silvana Pinheiro: conceptualization, writing – review & editing

Conflicts of interest

There are no conflicts to declare.

Supporting Information

The authors have cited additional references within the Supporting Information.[143-146]

Acknowledgments

The authors thank the Vice Chancellor for Research of the University of Costa Rica for its support work via the research projects 115-C2-126 and 908-C3-610. We also thank Dr. Antonio Viayna of the University of Barcelona for his comments for the improvement of this manuscript.

References

[1] M. J. Waring, *Expert Opin. Drug Discov.* **2010**, *5*, 235–248.

[2] D. F. V. Lewis, M. N. Jacobs, M. Dickins, *Drug Discov. Today* **2004**, *9*, 530–537.

[3] M. Chatzopoulou, E. Emer, C. Lecci, J. A. Rowley, A. S. Casagrande, L. Moir, S. E. Squire, S. G. Davies, S. Harriman, G. M. Wynne, F. X. Wilson, K. E. Davies, A. J. Russell, *ACS Med. Chem. Lett.* **2020**, *11*, 2421–2427.

[4] M. M. Miller, S. P. Wasik, G. L. Huang, W. Y. Shlu, D. Mackay, *Environ. Sci. Technol.* **1985**, *19*, 522–529.

[5] K. Soliman, F. Grimm, C. A. Wurm, A. Egner, *Sci. Rep.* **2021**, *11*, 1–9.

[6] H. O. Esser, *Pestic. Sci.* **1986**, *17*, 265–276.

[7] C. A. Lipinski, F. Lombardo, B. W. Dominy, P. J. Feeney, *Adv. Drug. Deliv. Rev.* **2001**, *46*, 3–26.

[8] D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward, K. D. Kopple, *J. Med. Chem.* **2002**, *45*, 2615–2623.

[9] A. Leo, C. Hansch, C. Church, *J. Med. Chem.* **1969**, *12*, 766–771.

[10] A. Leo, C. Hansch, D. Elkins, *Chem. Rev.* **1971**, *71*, 525–616.

[11] E. H. Kerns, *J. Pharm. Sci* **2001**, *90*, 1838–1858.

[12] X. Liu, M. Tu, R. S. Kelly, C. Chen, B. J. Smith, *Drug Metab. Dispos.* **2004**, *32*, 132–139.

[13] G. Colmenarejo, *Med. Res. Rev.* **2003**, *23*, 275–301.

[14] W. J. Zamora, C. Curutchet, J. M. Campanera, F. J. Luque, *J. Phys. Chem. B* **2017**, *121*, 9868–9880.

[15] V. Iglesias, C. Pintado-Grima, J. Santos, M. Fornt, S. Ventura, in *Data Mining Techniques for the Life Sciences* (Eds.: O. Carugo, F. Eisenhaber), Springer US, New York, NY, **2022**, pp. 197–211.

[16] M. Oeller, R. Kang, R. Bell, H. Ausserwöger, P. Sormanni, M. Vendruscolo, *Brief. Bioinform.* **2023**, *24*, bbad004.

[17] W. F. Porto, K. C. V. Ferreira, S. M. Ribeiro, O. L. Franco, *Biochim Biophys Acta Gen Subj* **2022**, *1866*, 130070.

[18] W. J. Zamora, S. De Souza, F. Separovic, Fco. J. Luque, *Biophys. J.* **2020**, *118*, 236a.

[19] S. Simm, J. Einloft, O. Mirus, E. Schleiff, *Biol.Res.* **2016**, *49*, 1–19.

[20] W. J. Zamora, J. M. Campanera, F. J. Luque, *J. Phys. Chem. Lett.* **2019**, *10*, 883–889.

[21] T. Ingram, U. Richter, T. Mehling, I. Smirnova, *Fluid Ph. Equilibria* **2011**, *305*, 197–203.

[22] C.-S. Chen, S.-T. Lin, *Ind. Eng. Chem. Res.* **2016**, *55*, 9284–9294.

[23] J. C. Westall, C. Leuenberger, R. P. Schwarzenbach, *Environ. Sci. Technol.* **1985**, *19*, 193–198.

[24] L. Xing, R. C. Glen, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796–805.

[25] I. V. Tetko, G. I. Poda, *J. Med. Chem.* **2004**, *47*, 5601–5604.

[26] D. Livingstone, *Curr. Top. Med. Chem.* **2005**, *3*, 1171–1192.

[27] C. C. Bannan, G. Calabró, D. Y. Kyu, D. L. Mobley, *J. Chem. Theory Comput.* **2016**, *12*, 4015–4024.

[28] T. D. Bergazin, N. Tielker, Y. Zhang, J. Mao, M. R. Gunner, K. Francisco, C. Ballatore, S. M. Kast, D. L. Mobley, *J. Comput. Aided. Mol. Des.* **2021**, *35*, 771–802.

[29] A. Avdeef, *Absorption and Drug Development.*, John Wiley & Sons, **2012**.

[30] A. Port, M. Bordas, R. Enrech, R. Pascual, M. Rosés, C. Ràfols, X. Subirats, E. Bosch, *Eur. J. Pharm. Sci.* **2018**, *122*, 331–340.

[31] R. P. Austin, P. Barton, A. M. Davis, C. N. Manners, M. C. Stansfield, *J. Pharm. Sci* **1998**, *87*, 599–607.

[32] P. Jain, A. Kumar, *Phys. Chem. Chem. Phys.* **2015**, *18*, 1105–1113.

[33] C. T. Jafvert, J. C. Westall, E. Grieder, R. P. Schwarzenbach, *Environ. Sci. Technol.* **1990**, *24*, 1795–1803.

[34] R. P. Austin, A. M. Davis, C. N. Manners, *J. Pharm. Sci* **1995**, *84*, 1180–1183.

[35] K. Takács-Novák, G. Szász, *Pharm. Res.* **1999**, *16*, 1633–1638.

[36] A. Fini, G. Fazio, M. Gonzalez-Rodriguez, C. Cavallari, N. Passerini, L. Rodriguez, *Int. J. Pharm.* **1999**, *187*, 163–173.

[37] V. Sarveiya, J. F. Templeton, H. A. E. Benson, *J. Pharm. Pharmacol.* **2004**, *56*, 717–724.

[38] R. A. Scherrer, S. F. Donovan, *Anal. Chem.* **2009**, *81*, 2768–2778.

[39] M. C. Wenlock, T. Potter, P. Barton, R. P. Austin, *SLAS Discov.* **2011**, *16*, 348–355.

[40] A. Fini, G. Bassini, A. Monastero, C. Cavallari, *Pharmaceutics* **2012**, *4*, 413–429.

[41] M. Paternostre, O. Meyer, C. Grabielle-Madelmont, S. Lesieur, M. Ghanam, M. Ollivon, *Biophys. J.* **1995**, *69*, 2476–2488.

[42] T. Pieńko, M. Grudzień, P. P. Taciak, A. P. Mazurek, *J. Mol. Graph. Model.* **2016**, *63*, 15–21.

[43] L. Quoc Hung, *J. Electroanal. Chem. Interfacial electrochem.* **1980**, *115*, 159–174.

[44] T. Kakiuchi, *Anal. Chem.* **1996**, *68*, 3658–3664.

[45] Z. Disdier, S. Savoye, R. V. H. Dagnelie, *Chemosphere* **2022**, *304*, DOI 10.1016/j.chemosphere.2022.135155.

[46] A. Berthod, S. Carda-Broch, M. C. Garcia-Alvarez-Coque, *Anal. Chem.* **1999**, *71*, 879–888.

[47] F. Âde, R. Reymond, P.-A. Carrupt, B. Testa, H. H. Girault, *Chem. Eur. J.* **1999**, *5*, 39–48.

[48] V. Gobry, S. Ulmeanu, F. Reymond, G. Bouchard, P. A. Carrupt, B. Testa, H. H. Girault, *J. Am. Chem. Soc.* **2001**, *123*, 10684–10690.

[49] F. Reymond, G. Steyaert, P. A. Carrupt, B. Testa, H. Girault, *J. Am. Chem. Soc.* **1996**, *118*, 11951–11957.

[50] R. D. Cunha, L. J. Ferreira, E. Orestes, M. D. Coutinho-Neto, J. M. de Almeida, R. M. Carvalho, C. D. Maciel, C. Curutchet, P. Homem-de-Mello, *Computation* **2022**, *10*, 170.

[51] S. Tshepelevitsh, K. Hernits, I. Leito, *J. Comput. Aided. Mol. Des.* **2018**, *32*, 711–722.

[52] S. Losada-Barreiro, F. Paiva-Martins, C. Bravo-Díaz, *Antioxidants* **2023**, *12*, 828.

[53] B. I. Escher, R. Abagyan, M. Embry, N. Klüver, A. D. Redman, C. Zarfl, T. F. Parkerton, *Environ. Toxicol. Chem.* **2020**, *39*, 269–286.

[54] M. A. C. K. Hansima, F. Zvomuya, I. Amarakoon, *Sci. Total Environ.* **2023**, *892*, 164387.

[55] A. Tsantili-Kakoulidou, I. Panderi, F. Csizmadia, F. Darvas, *J.Am. Pharm. Assoc.* **1997**, *86*, 1173–1179.

[56] W. J. Zamora, E. Bertsch, S. Suñer, S. Pinheiro, **2023**, DOI 10.5281/ZENODO.7956685.

[57] 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, *Nucleic Acids Res.* **2021**, *49*, D1388–D1395.

[58] B. Mishra, C. Sankar, M. Mishra, *J. Drug Target.* **2011**, *19*, 204–211.

[59] M. L. Bello, A. M. Junior, C. A. Freitas, M. L. A. Moreira, J. P. da Costa, M. A. de Souza, B. A. M. C. Santos, V. P. de Sousa, H. C. Castro, C. R. Rodrigues, L. M. Cabral, *Eur. J. Pharm. Sci.* **2022**, *175*, DOI 10.1016/j.ejps.2022.106222.

[60] J. Bezençon, M. B. Wittwer, B. Cutting, M. Smieško, B. Wagner, M. Kansy, B. Ernst, *J. Pharm. Biomed. Anal.* **2014**, *93*, 147–155.

[61] W. Voigt, R. Mannhold, J. Limberg, G. Blaschke, *J. Pharm. Sci* **1988**, *77*, 1018–1020.

[62] T. J. Roseman', S. H. Yalkowsky, *J. Pharm. Sci* **1973**, *62*, 1680–1685.

[63] M. Shalaeva, J. Kenseth, F. Lombardo, A. Bastin, *J. Pharm. Sci* **2008**, *97*, 2581–2606.

[64] Z. Qiang, C. Adams, *Water Res.* **2004**, *38*, 2874–2890.

[65] T. de A. D. dos Santos, D. O. da Costa, S. S. da R. Pita, F. S. Semaan, *Ecl. Quím* **2010**, *35*, 81–86.

[66] K. Morimotol, A. Nagayasu, S. Fukanoki, K. Morisaka, S.-H. Hyon2, Y. Ikada, *Drug. dev. Ind. Pharm.* **1990**, *16*, 13–29.

[67] T. Loftsson, S. Thorisdóttir, H. Fridriksdóttir, E. Stefánsson, *Acta Ophthalmol.* **2010**, *88*, 337–341.

[68] R. Mannhold, K. P. Dross, R. FRekker, van der Steen, *Quant. Stuct-Act. Relat* **1990**, *9*, 21–28.

[69] T. Loftsson, S. B. Vogensen, C. Desbos, P. Jansook, *AAPS PharmSciTech* **2008**, *9*, 425–430.

[70] N. Kuntworbe, R. G. Alany, M. Brimble, R. Al-Kassas, *Pharm. Dev. Technol.* **2013**, *18*, 866–876.

[71] M. S. Islam, M. M. Narurkar, *J. Pharm. Pharmacol.* **1993**, *45*, 682–686.

[72] Y. Deng, B. Li, K. Yu, T. Zhang, *Sci. Total Environ.* **2016**, *544*, 980–986.

[73] U. Franke, A. Munk, M. Wiese, *J. Pharm. Sci* **1999**, *88*, 89–95.

[74] A. Avdeef, K. J. Box, J. E. A. Comer, C. Hibbert, K. Y. Tam, *Pharm. Res.* **1998**, *15*, 209–215.

[75] R. H. K. Thanacoody, *Recent Pat. Antiinfect. Drug. Discov.* **2011**, *6*, 92–98.

[76] V. Martínez, M. I. Maguregui, R. M. Jiménez, R. M. Alonso, *J. Pharm. Biomed. Anal.* **2000**, *23*, 459–468.

[77] B. Huerta, A. Jakimska, M. Gros, S. Rodríguez-Mozaz, D. Barceló, *J. Chromatogr. A.* **2013**, *1288*, 63–72.

[78] A. Fini, G. Fazio, G. Feroci, *Int. J. Pharm.* **1995**, *126*, 95–102.

[79] M. R. Jacka, *Clarke's Isolation and Identification of Drugs*, Pharmaceutical Press, **2000**.

[80] Y. Nakamura, H. Yamamoto, J. Sekizawa, T. Kondo, N. Hirai, N. Tatarazako, *Chemosphere* **2008**, *70*, 865–873.

[81] W. Schräder, J. T. Andersson, *J. Pharm. Sci* **2001**, *90*, 1948–1954.

[82] A. Avdeef, *Sirius Technical Application Notes (STAN)*, Sirius Analytical Instruments Ltd., **1994**.

[83] G. Caron, G. Steyaert, A. Pagliara, F. Âde, Â. Reymond, P. Crivori, P. Gaillard, P.-A. Carrupt, A. Avdeef, J. Comer, K. J. Box, H. H. Girault, B. Testa, **n.d.**, DOI 10.1002/(SICI)1522-2675(19990804)82:8.

[84] A. Avdeef, *Sirius Technical Application Notes (STAN)*, Sirius Analytical Instruments Ltd., **1995**.

[85] F. Lombardo, M. Y. Shalaeva, K. A. Tupper, F. Gao, M. H. Abraham, *J. Med. Chem.* **2000**, *43*, 2922–2928.

[86] S. Winiwarter, N. M. Bonham, F. Ax, A. Hallberg, H. Lennernäs, A. Karlén, *J. Med. Chem.* **1998**, *41*, 4939–4949.

[87] B. Slater, A. McCormack, A. Avdeef, J. E. A. Comer, *J. Pharm. Sci* **1994**, *83*, 1280–1283.

[88] P. Luger, K. Daneck, W. Engel, G. Trummlitz, K. Wagner, *Eur. J. Pharm. Sci.* **1996**, *4*, 175–187.

[89] K. Takács-Novák, M. Józan, I. Hermecz, G. Szász, *Int. J. Pharm.* **1992**, *79*, 89–96.

[90] S. Carda-Broch, A. Berthod, *J. Chromatogr. A.* **2003**, *995*, 55–66.

[91] D. Scott, *Pharm. Technol.* **2002**, *26*.

[92] R. Gulaboski, F. Borges, C. M. Pereira, M. Natália, D. S. Cordeiro, J. Garrido, A. F. Silva, *Comb. Chem. High Throughput Screen.* **2007**, *10*, 514–526.

[93] "pKa Plugin | Chemaxon Docs," can be found under https://docs.chemaxon.com/display/docs/pka-plugin.md, **n.d.**

[94] "LogP and logD calculations | Chemaxon Docs," can be found under https://docs.chemaxon.com/display/docs/logp-and-logd-calculations.md, **n.d.**

[95] A. K. Ghose, G. M. Crippen, *J. Comput. Chem.* **1986**, *7*, 565–577.

[96] B. Hammer, M. Franco, E. LeDell, "Package 'Metrics,'" can be found under https://cran.r-project.org/web/packages/Metrics/Metrics.pdf, **2022**.

[97] J. Freiría-Gándara, S. Losada-Barreiro, F. Paiva-Martins, C. Bravo-Díaz, *J. Chem. Eng. Data* **2018**, *63*, 2999–3007.

[98] N. W. Shappell, W. L. Shelver, S. J. Lupton, W. Fanaselle, J. M. Van Doren, H. Hakk, *J. Agric. Food Chem.* **2017**, *65*, 938–949.

[99] S. J. Lupton, N. W. Shappell, W. L. Shelver, H. Hakk, *J. Agric. Food Chem.* **2018**, *66*, 306–314.

[100] J. L. Colaizzi, P. R. Klink, *J. Pharm. Sci.* **1969**, *58*, 1184–1189.

[101] L. Niu, L. Henneberger, J. Huchthausen, M. Krauss, A. Ogefere, B. I. Escher, *ACS Environ. Au* **2022**, *2*, 253–262.

[102] P. D. Leeson, B. Springthorpe, *Nat Rev Drug Discov* **2007**, *6*, 881–890.

[103] J. S. Scott, M. J. Waring, *Bioorg. Med. Chem.* **2018**, *26*, 3006–3015.

[104] T. W. Johnson, P. F. Richardson, S. Bailey, A. Brooun, B. J. Burke, M. R. Collins, J. J. Cui, J. G. Deal, Y.-L. Deng, D. Dinh, L. D. Engstrom, M. He, J. Hoffman, R. L. Hoffman, Q. Huang, R. S. Kania, J. C. Kath, H. Lam, J. L. Lam, P. T. Le, L. Lingardo, W. Liu, M. McTigue, C. L. Palmer, N. W. Sach, T. Smeal, G. L. Smith, A. E. Stewart, S. Timofeevski, H. Zhu, J. Zhu, H. Y. Zou, M. P. Edwards, *J. Med. Chem.* **2014**, *57*, 4720–4744.

[105] J. J. Cui, M. Tran-Dubé, H. Shen, M. Nambu, P.-P. Kung, M. Pairish, L. Jia, J. Meng, L. Funk, I. Botrous, M. McTigue, N. Grodsky, K. Ryan, E. Padrique, G. Alton, S. Timofeevski, S. Yamazaki, Q. Li, H. Zou, J. Christensen, B. Mroczkowski, S. Bender, R. S. Kania, M. P. Edwards, *J. Med. Chem.* **2011**, *54*, 6342–6363.

[106] L. Z. Benet, F. Broccatelli, T. I. Oprea, *AAPS J.* **2011**, *13*, 519–547.

[107] P. Di Fruscia, F. Edfeldt, I. Shamovsky, G. W. Collie, A. Aagaard, L. Barlind, U. Börjesson, E. L. Hansson, R. J. Lewis, M. K. Nilsson, L. Öster, J. Pemberton, L. Ripa, R. I. Storer, H. Käck, *ACS Med. Chem. Lett.* **2021**, *12*, 302–308.

[108] M. F. Sammons, S. V. Kharade, K. J. Filipski, M. Boehm, A. C. Smith, A. Shavnya, D. P. Fernando, M. S. Dowling, P. A. Carpino, N. A. Castle, S. G. Zellmer, B. M. Antonio, J. R. Gosset, A. Carlo, J. S. Denton, *ACS Med. Chem. Lett.* **2018**, *9*, 125–130.

[109] K. Hoegenauer, J. Kallen, E. Jiménez-Núñez, R. Strang, P. Ertl, N. G. Cooke, S. Hintermann, M. Voegtle, C. Betschart, D. J. J. McKay, J. Wagner, J. Ottl, C. Beerli, A. Billich, J. Dawson, K. Kaupmann, M. Streiff, N. Gobeau, S. Harlfinger, R. Stringer, C. Guntermann, *J. Med. Chem.* **2019**, *62*, 10816–10832.

[110] Á. Tarcsay, K. Nyíri, G. M. Keserű, *J. Med. Chem.* **2012**, *55*, 1252–1260.

[111] F. McTaggart, L. Buckett, R. Davidson, G. Holdgate, A. McCormick, D. Schneck, G. Smith, M. Warwick, *Am. J. Cardiol.* **2001**, *87*, 28–32.

[112] C. M. Bowman, F. Ma, J. Mao, Y. Chen, *CPT: Pharmacomet. Syst. Pharmacol.* **2021**, *10*, 5–17.

[113] T. W. Johnson, S. P. Tanis, S. L. Butler, D. Dalvie, D. M. DeLisle, K. R. Dress, E. J. Flahive, Q. Hu, J. E. Kuehler, A. Kuki, W. Liu, G. A. McClellan, Q. Peng, M. B. Plewe, P. F. Richardson, G. L. Smith, J. Solowiej, K. T. Tran, H. Wang, X. Yu, J. Zhang, H. Zhu, *J. Med. Chem.* **2011**, *54*, 3393–3417.

[114] S. T. M. Orr, R. Beveridge, S. K. Bhattacharya, K. O. Cameron, S. Coffey, D. Fernando, D. Hepworth, M. V. Jackson, V. Khot, R. Kosa, K. Lapham, P. M. Loria, K. F. McClure, J. Patel, C. Rose, J. Saenz, I. A. Stock, G. Storer, M. von Volkenburg, D. Vrieze, G. Wang, J. Xiao, Y. Zhang, *ACS Med. Chem. Lett.* **2015**, *6*, 156–161.

[115] E. L. Willighagen, *Groovy Cheminformatics with the Chemistry Development Kit*, **2023**.

[116] G. M. Ghiandoni, E. Caldeweyher, *Sci. Rep.* **2023**, *13*, 1–11.

[117] B. L. Welch, *Biometrika* **1947**, *34*, 28–35.

[118] G. D. Ruxton, *Behav. Ecol.* **2006**, *17*, 688–690.

[119] M. Kuhn, K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*, Taylor And Francis Group, **2019**.

[120] J. Wing, S. Weston, A. Williams, C. Keefer, A. Ebgelhardt, T. Cooper, Z. Mayer, B. Kenkel, "Package 'caret,'" can be found under https://cran.r-project.org/web/packages/caret/caret.pdf, **2023**.

[121] S. Liu, C. Cao, Z. Li, *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387–394.

[122] L. H. Hall, L. B. Kier, *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.

[123] A. I. McLeod, C. Xu, Y. Lai, "Package 'bestglm,'" can be found under https://cran.r-project.org/web/packages/bestglm/bestglm.pdf, **2022**.

[124] G. M. Furnival, R. W. Wilson, *Technometrics* **1974**, *16*, 499–511.

[125] S. Burger, *Introduction to Machine Learning with R: Rigorous Mathematical Modeling*, O'Reilly, **2018**.

[126] J. Fauchere, V. Pliska, *Eur. J. Med. Chem.* **1983**, *18*.

[127] L. Breiman, *Mach. Learn.* **2001**, *45*, 5–32.

[128] L. Breiman, A. Cutler, A. Liaw, M. Wiener, "Package 'randomForest,'" can be found under https://cran.r-project.org/web/packages/randomForest/randomForest.pdf, **2022**.

[129] C. Cortes, V. Vapnik, L. Saitta, *Mach. Learn.* **1995**, *20*, 273–297.

[130] B. E. Boser, I. M. Guyon, V. N. Vapnik, *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory* **1992**, 144–152.

[131] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, C.-C. Chang, C.-C. Lin, "Package 'e1071,'" can be found under https://cran.r-project.org/web/packages/e1071/e1071.pdf, **2022**.

[132] S. Kundu, S. J. Perinjelil, N. Thakur, in *Mitigation of Plant Abiotic Stress by Microorganisms* (Eds.: G. Santoyo, A. Kumar, M. Aamir, S. Uthandi), Academic Press, **2022**, pp. 231–256.

[133] Institute of Medicine (US) Committee on Strategies toReduce Sodium Intake, *Strategies to Reduce Sodium Intake in the United States*, National Academies Press (US), Washington (DC), **2010**.

[134] K. Xing, S. Chatterjee, T. Saito, C. Gainaru, A. P. Sokolov, *Macromolecules* **2016**, *49*, 3138–3147.

[135] A. Viayna, S. G. Antermite, M. de Candia, C. D. Altomare, F. J. Luque, *J. Phys. Chem. B* **2020**, *124*, 28–37.

[136] N. Tielker, S. Güssregen, S. M. Kast, *J. Comput. Aided. Mol. Des.* **2021**, *35*, 933–941.

[137] A. Viayna, S. Pinheiro, C. Curutchet, F. J. Luque, W. J. Zamora, *J. Comput. Aided. Mol. Des.* **2021**, *35*, 803–811.

[138] S. A. Rodriguez, J. V. Tran, S. J. Sabatino, A. S. Paluch, *J. Comput. Aided. Mol. Des.* **2022**, *36*, 687–705.

[139] J. Wu, Y. Kang, P. Pan, T. Hou, *Drug Discov. Today* **2022**, *27*, 103372.

[140] R. C. Johnston, K. Yao, Z. Kaplan, M. Chelliah, K. Leswing, S. Seekins, S. Watts, D. Calkins, J. Chief Elk, S. V. Jerome, M. P. Repasky, J. C. Shelley, *J. Chem. Theory Comput.* **2023**, *19*, 2380–2388.

[141] N. Šegatin, C. Klofutar, *Monatsch. Für. Chem.* **2004**, *135*, 241–248.

[142] F. Reymond, V. Chopineaux-Courtois, G. Steyaert, G. Bouchard, P. A. Carrupt, B. Testa, H. H. Girault, *J. Electroanal. Chem.* **1999**, *462*, 235–250.

[143] F. R. Burden, *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 225–227.

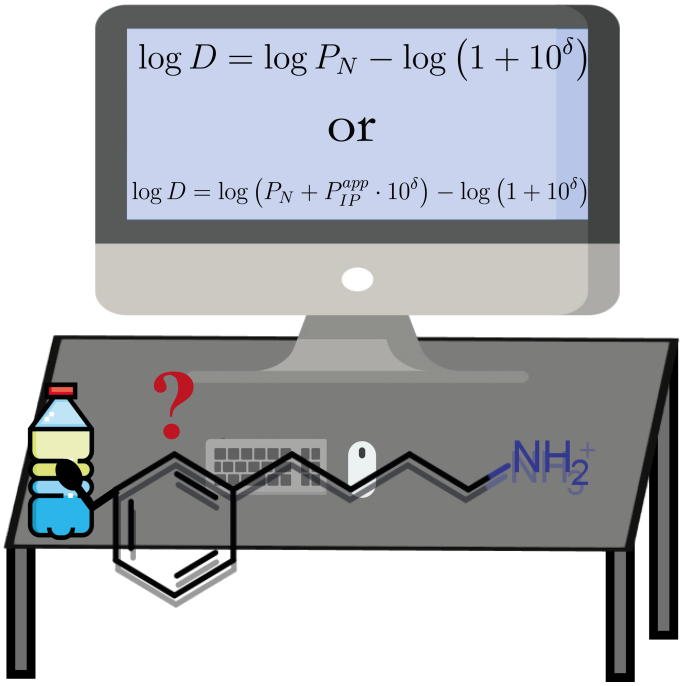
[144] F. R. Burden, *Quant. Stuct-Act. Relat* **1997**, *16*, 3–314.

[145] R. S. Pearlman, K. M. Smith, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28–35.

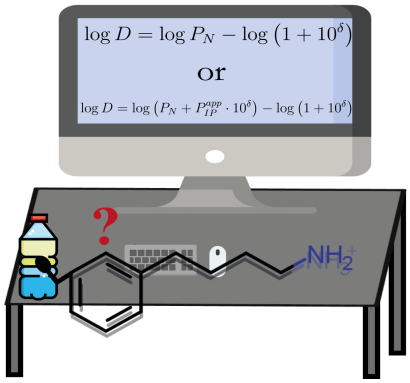
[146] M. Petitjean, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 331–337.

10.1002/cphc.202300548

Table of Contents



Lipophilicity is a property with wide relevance in drug design, computational biology, food, environmental and medicinal chemistry. Herein, a 4-phenylbutylamine molecule is searching in a computer for which lipophilicity profile will give a more accurate prediction for its distribution coefficient. We demonstrated that considering the apparent ion pair partitioning gives more accurate results.



Lipophilicity is a property with wide relevance in drug design, computational biology, food, environmental and medicinal chemistry. Herein, a 4-phenylbutylamine molecule is searching in a computer for which lipophilicity profile will give a more accurate prediction for its distribution coefficient. We demonstrated that considering the apparent ion pair partitioning gives more accurate results.

10.1002/cphc.202300548

Twitter

Author	Twitter handle
Esteban Bertsch	@bertschito
Sebastián Suñer	-
Dr. Sylvana Pinheiro	-
Prof. Dr. William J. Zamora	@willzamorachem