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

Esteban Bertsch[1+], Sebastian Suñer[1+], Silvana Pinheiro[1,2], 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 (WJZ)

[+]These authors contributed equally

Keywords Partition coefficient, Lipophilic profiles, Machine learning, Chemoinformatics, Drug Design, pH.

Abstract

Lipophilicity is a physicochemical property with wide relevance in drug design and also applied in areas such as food chemistry, environmental chemistry, and computational biology. This descriptor strongly influences the absorption, distribution, permeability, bioaccumulation, protein-binding, and biological activity of bioorganic compounds. Lipophilicity is commonly expressed as the *n*-octanol/water partition coefficient (P_N) 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 ionic partition (P_I^{app}) because of the challenge of predicting the partitioning of the charged species and/or related species (e.g., ion-pairs, counterions, molecular aggregates). In this work, we studied the impact of P_I^{app} on the prediction of lipophilicity of small molecules by modeling 225 logD_{pH} of a set of experimental values using the formalism that takes into account a pH correction (see Eq. 1) and the one considering the apparent partition of ionic species (see Eq. 2). Our findings show that a better fit is obtained by considering the apparent ionic partition while ignoring its contribution can lead to inadequate computational predictions. In this context, we developed machine learning algorithms to determine in which cases the 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 Eq. 2. In addition, we validated our findings using a test and two external set which include small molecules and amino acids analogs where the logistic regressions, random forest classifications, and support vector machine models predicted the better formalism to determine the logD_{pH} for each molecule with high accuracies, sensitivities, and specificities. 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 pH-dependent lipophilicity profile should be used depending on the structure of a substance in their research.

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 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 a more proper descriptor for human bioavailability due to the frequent pH-dependence of drugs. This property has shown be useful in QSAR models to explain how small molecules have human brain cells permeability[12] or binding 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 the experimentally determined logD_{pH} values, theoretical lipophilicity profiles give 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 a mass balance between 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.[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 displays the equilibria from which it is derived. Eq. 1 has been used to easily calculate $\log D_{pH}$ from $\log P_N$ values obtained by empirical computational models.[24-26] This equation was widely used in $\log D_{pH}$ estimation methods in the SAMPL6 and SAMPL7 blind challenge, which is a large-scale comparative evaluation for drug design predictive models.[27,28]

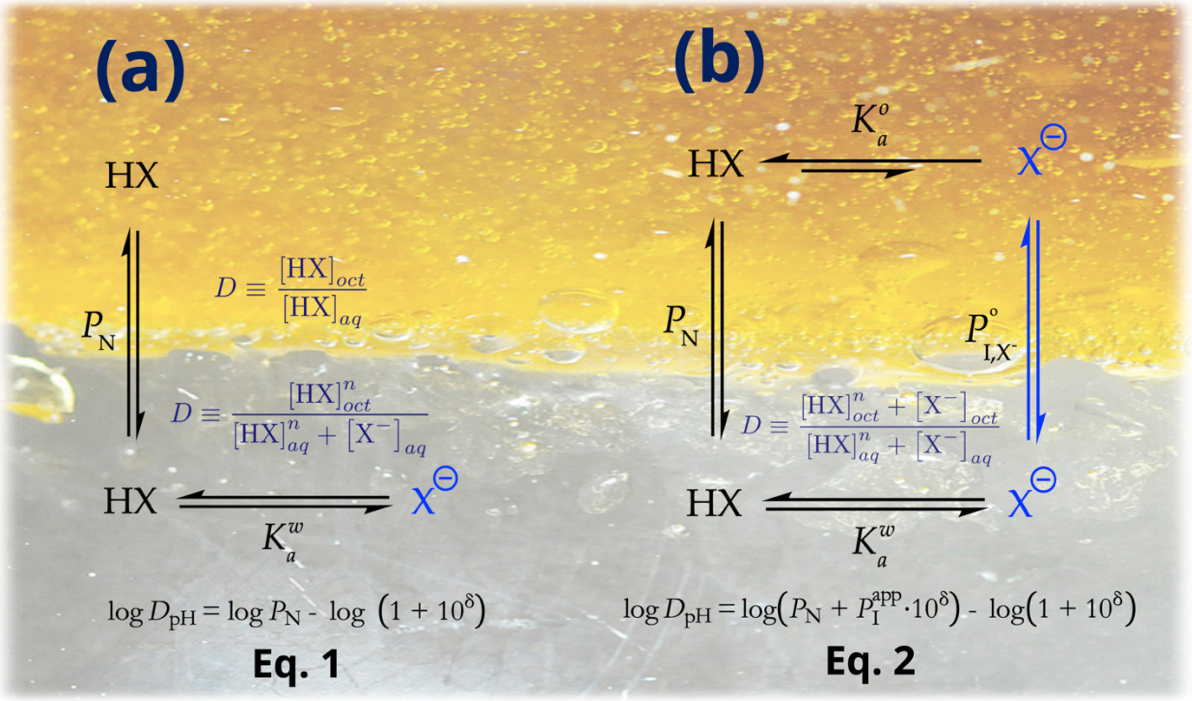


Figure 1. Representations of the mechanism of partition for a symbolic ionizable acidic molecule for both neutral (HX) and ionic (X−) species using (a) Equation 1 and (b) Equation 2. The theoretical partition of the charged organic specie (P_{I,X^-}) has been replaced by experimental measurable apparent partitioning (P_I^{app}) in Eq. 2.

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

$$\log D_{pH} = \log\left(P_N + P_I^{app} \cdot 10^{\delta}\right) - \log\left(1 + 10^{\delta}\right) \quad [2]$$

Equation 2 is commonly called ionic partition P_I model [29], which represents a simplification that only considers the partition of the charged organic specie (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 specie (es) involved in the partitioning, hence, the partition of ionic species is measured as an apparent partitioning (P_I^{app}). This experimentally measurable apparent partition coefficient depends on the background salt[31], compound concentration[32], and may involve much more complex species such as ion-pairs[33-40], and aggregates[41]. Some studies have simplified the P_I^{app} to the partition of only ionic organic species (P_I) because these methods used have been parametrized by using experimental P_I^{app} values[14,42], while other theoretical studies have modeled it using the participation of ion-pairs (P_{IP})[21,22]. Recently, an alternative model[14] to that of 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_I^{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 they fitted their 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_I^{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 the P_I^{app} and considering that different theoretical approaches have shown similar trends[14,21,22], Equation 2 has been successfully used for modeling lipophilicity of ionized compounds in many areas of basic and applied sciences. For instance, to study 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]

Previous studies have 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] It was reported that Equation 1 tends to overestimate the hydrophobicity of the studied molecules, given that the P_I^{app} is not considered, on the other hand, Equation 2 predicts a logD_{pH} value closer to the experimental values. This study showed that Equation 2 provides a more exact lipophilicity profile at a wider pH range than Equation 1. However, no systematized study has been performed to evaluate the importance of considering the ionic partition on the logD_{pH} prediction for large sets of small drug-like molecules at various pH values although when it has been reported that much of the poor performance of some models on blind remains has been due to the simplification of ignoring the ionic species partition.[27]

In this study, we aim to provide 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 pH-dependent lipophilicity profile should be used depending on the structure of a substance in their research. For this, we collected the experimental values of logP_N, pK_a, and logP_I^{app} of different compounds at various pH values, which are used to compute logD_{pH} with Equation 1 and Equation 2. We compared both calculations through statistical parameters with the experimental logD_{pH} values. In addition, logistic regression (**LR**), random forest classification (**RFC**), and support vector machine (**SVM**) models are developed to define from the molecular structure which formalism is recommended for modeling a pH-depended lipophilicity profile.

Methodology

Data collection and classification

We critically compiled experimental values of logP_N, pK_a, logP_I^{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 logP_N, logD_{pH} and logP_I

values and because they accommodate the desired chemical space of small molecules for our modeling. SMILES codes were collected from publicly available data in PubChem[57] The pK_a values were also obtained from PubChem but they were also corroborated by reviewing their values in primary literature reports.[38,57-80] The experimental technique of logP_N, logD_{pH}, and logP_I^{app} measurements for each entry was thoroughly revised and added to the database.[81-91] Ref 55 provided experimental logD_{pH} values of molecules at diverse pH ranges. The logP_I values were obtained from the logD_{pH} at the most extreme measured pH, in which the molecule will be mostly (above 95 %) in its ionized state. The logP_I^{app} values for molecules that were not measured under ionizable pH conditions were obtained from external sources.[38,74,92,93] The molecules were classified as acids or bases based on their functional groups and pK_a values. Zwitterionic compounds were found by evaluating the difference between acidic and basic pK_a in conjunction with ChemAxon's calculator of protonated species distribution in function of pH.[94] Zwitterionic and amphoteric species were also classified as acidic or basic based on their behavior of their lipophilicity profiles, which were evaluated using the ChemAxon partitioning calculator.[95]

The experimental data for each molecule were used to compute the logD_{pH} values using Eq. 1 and Eq. 2 and are 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 performances of the two formalisms were 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 descriptor $\boldsymbol{d_3}$ (Eq. 5). When d_3 yields a value greater than zero, Eq. 2 fits a more appropriate lipophilicity value and vice versa.

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

We create 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 to fit equally well, which can be interpreted as Eq. 1 providing better modeling due to its simplicity.

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[97] while experimental descriptors ($\log P_{N,}$ $pK_{a,}$ and pH) were added from our dataset. We also added the free energies of hydration and hydrogen bond strengths computed using the open-source tool '*Jazzy*'[98] The H-bond donor and acceptor strengths were obtained by calculating the partial charges of 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 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.[98]

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 our Machine Learning models. Firstly, we performed a Welch's *t*-test (**WTT**), which evaluates the statistical difference between the means of two populations that have unequal variances and sample sizes.[99,100] The algorithm calculates the mean of both groups from the binomial conditional for each descriptor. These values are evaluated using Equation 6.

$$t = \frac{\Delta\mu}{\delta_{\Delta\bar{x}}} \qquad [6]$$

where *t* stands for the statistic *t* in the Welch's t-test, $\Delta\mu$ 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.7):

$$\delta_{\Delta\bar{x}} = \sqrt{\left(\frac{s_1}{\sqrt{N_1}}\right)^2 + \left(\frac{s_2}{\sqrt{N_2}}\right)^2} \quad [7]$$

The WTT was performed for each descriptor using R where the *p*-value was extracted. Features that did not show statistical significance between the means ($p > 0.05$) were eliminated. Secondly, a 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 (see Figure S3). The least important descriptors are removed, and the model was re-iterated to achieve maximum performance.[101] This RFE algorithm was programmed using the '*caret'* package in R[102] and tuned via a 5-time reiterated *k*-fold cross validation ($k = 10$). Table 1 shows the selected descriptors with the WTT feature selection method for acids and bases, along with their definitions and target molecules. Table S1 shows the descriptors selected using the RFE method.

Table 1. List of the most influential structural descriptors[98,103-106] used for the logistic regression 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 3D structure directed 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[107] which evaluates through *n* iterations, which combination of descriptors gives the best fitted regression through the *leaps* algorithm.[108] 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 8).

$$\ln\left(\frac{f(x)}{1-f(x)}\right) = \sum_{i=1}^{n} c_i x_i + \mathrm{b} \qquad [8]$$

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. This feature selection process was performed separately for acids and bases because the descriptors have different behaviors for each type of molecule.

Figure 2 shows a flowchart of the modelling process. The dataset was divided into acids (113 entries) and bases (100 entries). Zwitterions (7 entries) were not considered for the Machine Learning predictions because of their small sample size and because further lipophilicity modeling can be performed for these molecules (see Results and Discussion section). Acids and bases were randomly sampled into training and test sets at a ratio of 80:20. Multiple logistic regressions were performed for the training sets based on previously collected descriptors. Predictive models were programmed using the '*caret'* package. Acids and bases were modeled separately and labeled as **Models A** and **B**, respectively (see Figure 2). 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.[109] 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 9). On the other

hand, 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 10). The accuracy indicates the overall performance in detecting false positives and false negatives (Equation 11).

$$Sensitivity = \frac{TP}{TP + FN} \quad [9]$$

$$Specificity = \frac{TN}{FP + TN} \quad [10]$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad [11]$$

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 were evaluated[110], consisting of 8 entries of histidine (basic amino acid) and 10 entries of tyrosine (acidic amino acid). Then, we evaluated the performance of Model A and Model B for this data set with confusion matrices (see Table S3-S4).

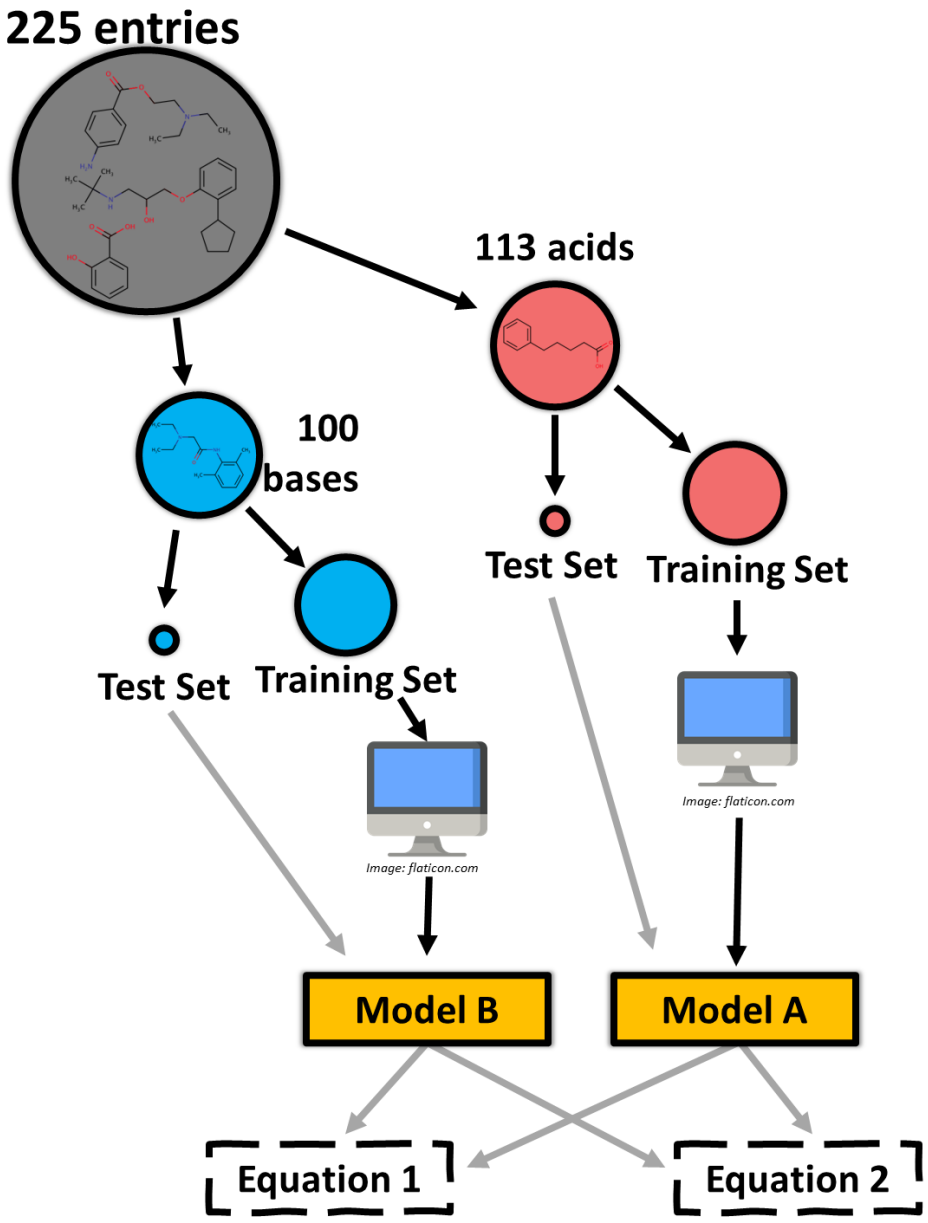


Figure 2. Graphical representation of the data classification and sampling of our dataset to create our predictive multiple logistic regression model using topological, constitutional, and experimental descriptors.

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.[111] We split our dataset, training-, and test-sets as shown in Figure 2. In this case, Model A and Model B consist of random forest classification (**RFC**) performed with the '*randomForest*' package in R.[112] 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 gave 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 (see 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. 9-11).

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.[113] When the data are not linearly separable, the algorithm performs the kernel trick, which consists of 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.[114]

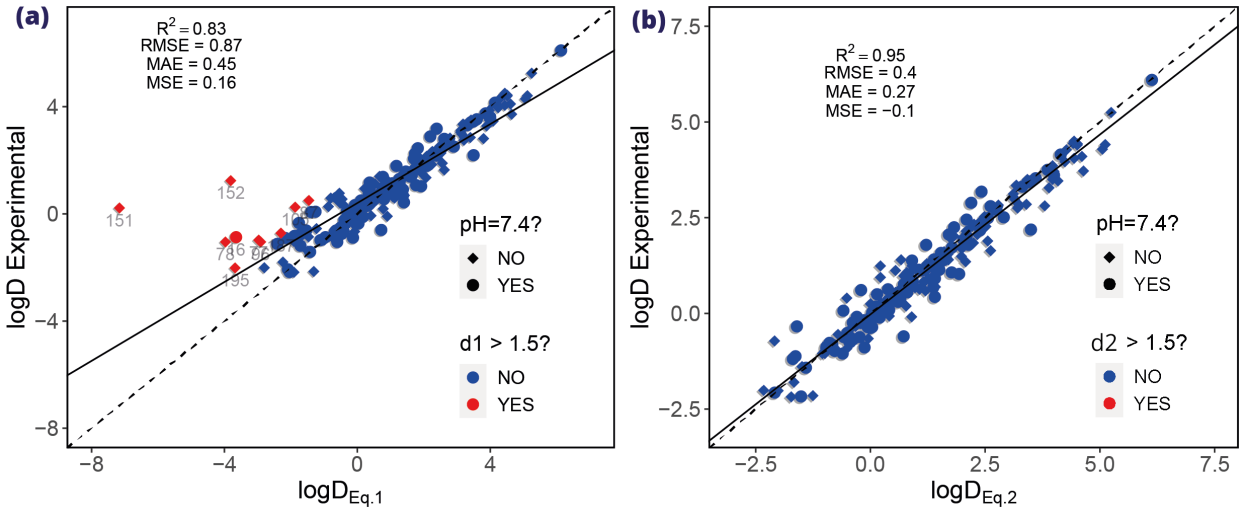
We split our datasets in the same manner as with the other classification models and set Model A and Model B as support vector machines given by the *'e1071'* package in R.[115] We decided to compare the performance of using a linear kernel (**SVML**) and a polynomial kernel (**SVMP**); radial kernels were not evaluated because our binary data do not follow a circular

separation by the hyperplane, so it does not adequately fit a radial kernel SVM classification. 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

Our database consists of pK_a, logP_N, logP_I^{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 and collaborators containing logD_{pH} values at various pH for each molecule as an individual entry.[55] Molecules with logD_{pH} values measured in the presence of background salt concentrations above 0.15 mol/L were discarded because the study of the effect of external ions on lipophilicity is beyond the scope of our study. Thus, we finally obtained 225 entries (118 individual molecules) with 113 acids, 100 bases, and 12 zwitterions.

Calculation of logD_{pH} was accomplished using Eq. 1 and Eq. 2 for each molecule at their respective pH. Figure 3 shows the overall performance of each model by comparing the computed 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 the Eq. 1 (see Figure 3a, red marks) with highly underestimated predictions. As a consequence, Eq. 1 poorly predicts the logD_{pH} values at extreme pH. On the other hand, the predicted values using the Eq. 2 are significantly better (see Figure 3b), reducing the RMSE by 0.48 log*D* units, which represents an improvement of 55% in accuracy.



355

356 **Figure 3**. Evaluation of the computed logD_{pH} of our database compared with the experimental
357 values with (a) Eq. 1 and (b) Eq. 2. Rhomboids represent logD_{pH} when the pH is different of 7.4.
358 Red dots and rhomboids highlight compounds with a deviation greater than 1.5 log$D$ units.
359 Statistical parameters were calculated using the '*Metrics'* package in R (R^2 = squared Pearson's
360 correlation coefficient, RMSE = root mean squared error, MAE = mean absolute errorand , MSE
361 = mean squared error).

362

363 Table 2 shows the reduction of RMSE in log$D$ units of each molecule type by using Eq.2
364 instead Eq.1. It is observed that any type of molecule shows a significant improvement in its
365 performance when its distribution coefficient is modeled with log$D_{Eq.2}$ (see Figure S5). Basic
366 molecules showed the greatest improvement as the deviation shown by logDEq.1 was greater than
367 one unit of RMSE in log$D$ units. Zwitterions also showed a significant improvement, even though
368 these molecules can have multiple ionic partition coefficients (cationic partitions P^+, and anionic
369 partitions P^-, and zwitterionic partitions Pz), which are not considered in the model log$D_{Eq.2}$. These
370 partitions can be added by considering both acidic and basic pK_a into the thermodynamic
371 equilibria.[45] Despite this, the implementation of just one of the two P_I^{app} did a significant
372 improvement in the lipophilic modelling of zwitterions.

373

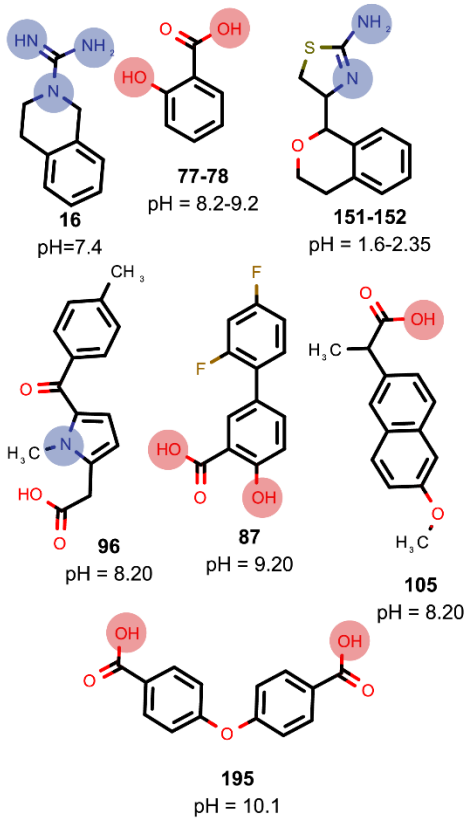
374

Table 2. Values of ΔRMSE 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 (Figure S1).

Type	ΔRMSE[a]
Acid	0.30
Base	0.67
Zwitterion	0.38
All	0.48

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

The molecules with the highest deviations in the prediction of the experimental $\log D_{pH}$ using the $\log D_{Eq1}$ are displayed in Figure 4. The chemical nature of the outliers is dominated by the presence of ionic species because these compounds were experimentally measured to extreme pH. These deviations respond to the theoretical framework of Eq. 1 and Eq. 2, thus, the inclusion of the term P_I^{app} in Eq. 2 corrects the prediction. Figure 4 shows various polyacids or amphoteric molecules with multiple ionizable sites included in our dataset. Bases **16, 151-152** have multiple protonation sites, while acids **77, 78, 87,** and **195** have two deprotonation sites, and amphoteric **96** has a carboxylic acid and a tertiary aromatic nitrogen that can protonate at certain pH values. The prediction of the $\log D_{pH}$ of these molecules can be improved by using more complex thermodynamic models considering several equilibria.[45] However, it is shown here that the consideration of one of the P_I^{app} with $\log D_{Eq.2}$ is enough to significantly increase the accuracy of the lipophilicity modeling of these compounds to extreme pH where one charged species can predominate over the others.



392

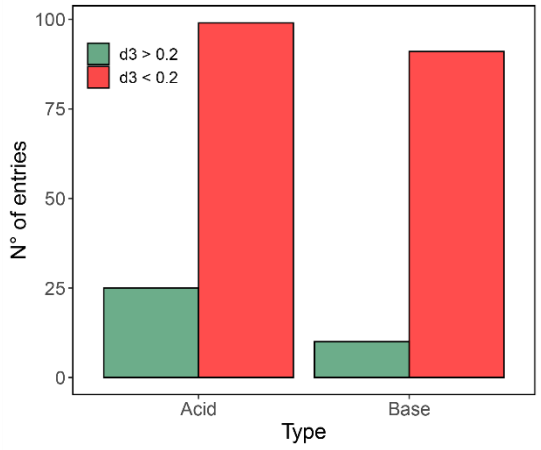
393 **Figure 4.** Representation of the molecules with the highest deviations in the prediction of the
394 experimental logD_{pH} using the logD_{Eq1}. The protonation and deprotonation sites of each molecule
395 were labeled in blue and red, respectively.

396

397 One of the aims of this study is to develop a classification algorithm that can differentiate
398 whether the lipophilicity profile of a molecule would be better predicted with log$D_{Eq.1}$ or log$D_{Eq.2}$.
399 However, we noticed that a significant number of entries yielded d_3 values close to 0 (see Figure
400 S6a), which denotes that both formalisms compute a similar result compared to their experimental
401 values. Therefore, let us note that we focus on the specific cases with a significant improvement
402 when the P_I^{app} of molecules is considered. Indeed, we decided to delimit the conditional d_3
403 indicating that if a molecule exceeds a certain value of d_3, it is important to consider its apparent
404 ionic partitioning for predicting its lipophilicity. We tested d_3 values between 0.1-1 and picked the
405 optimal value based on two parameters. Firstly, considering that our set is small due to the fact
406 that we used strictly experimental values in our database, we seek that the population of molecules
407 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 will have a larger number of parameters to create predictive models with higher accuracy. In consequence, the delimiter '0.2' showed an adequate balance between these two parameters, and it was selected as our cut-off value (see Figure S6b). Thus, 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 the $\log D_{\text{Eq.1}}$ due to its easy implementation (it does not depend on $P_{\text{I}}^{\text{app}}$, resulting in less computational effort and fewer experimental parameters). Higher thresholds significantly decreased the population in $\log D_{\text{Eq.2}}$, while lower values reduced the structural divergence between molecules in $\log D_{\text{Eq.1}}$ and $\log D_{\text{Eq.2}}$, making it more difficult to find structural descriptors that can differentiate between both populations. The value '0.5' was also tested since a local maximum of descriptors with p-values < 0.05 was observed at this point (see Figure S6b). Furthermore, this value is of experimental interest, because $\log P_{\text{N}}$ measurements of substances with different techniques tend to vary by amounts less than 0.5 log*P* units (using the Shake-Flask method as a reference), being this value considered as a parameter to indicate 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 the external set 1 (see Figure S7). This phenomenon can be explained since this d_3 delimiter has a very small $\log D_{\text{Eq.2}}$ 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 5 shows the distribution of the molecules in our database, classified using the criteria $d_3 > 0.2$ as binary descriptor. Most entries can be computed using $\log D_{\text{Eq.1}}$ 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 $\log D_{\text{Eq.2}}$.



436

437 **Figure 5**. Distribution of acid and basic entries from our dataset as a function of their d_3 values.

438

439 We obtained several structural and physicochemical descriptors of the molecules to find a
440 considerable divergence between the populations. First, our database was split into acids and bases
441 and then in training and test set. The '*rcdk'* package in R was used to look through the descriptors,
442 along with the *Jazzy* calculations of energies of hydration and hydrogen-bond strengths and the
443 experimental descriptors. The feature selection methods selected show a wide range of diverse
444 descriptors (see Table 1 and Table S1). Then, we performed a Welch's *t*-test on our descriptors
445 (WTT) where is analyzed the divergence between populations relative to the variances of the two
446 groups.[99] This test was selected over a Student's *t*-test because of the divergence of sample sizes
447 (Figure 5) and variances between groups (Figure 6-7).[100] The WTT descriptors gave acceptable
448 accuracies (see Figure S8).

449 An iterative feature selection method was also tested using an RFE model. The algorithm
450 achieved better performance when the 14 most important variables for acids and the nine most
451 important variables for bases were maintained. The importance of each descriptor posed by RFE
452 is shown in Figure S3. Good results were obtained when these descriptors were implemented in
453 the training of the machine learning models. However, the accuracy decreased significantly when
454 the test and external set 1 was evaluated (see Figure S8c-d), indicating that these descriptors did
455 not generate a sufficiently robust model, or a large number of chosen descriptors (see Table S1)
456 may overfit the data. Therefore, we selected the 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.

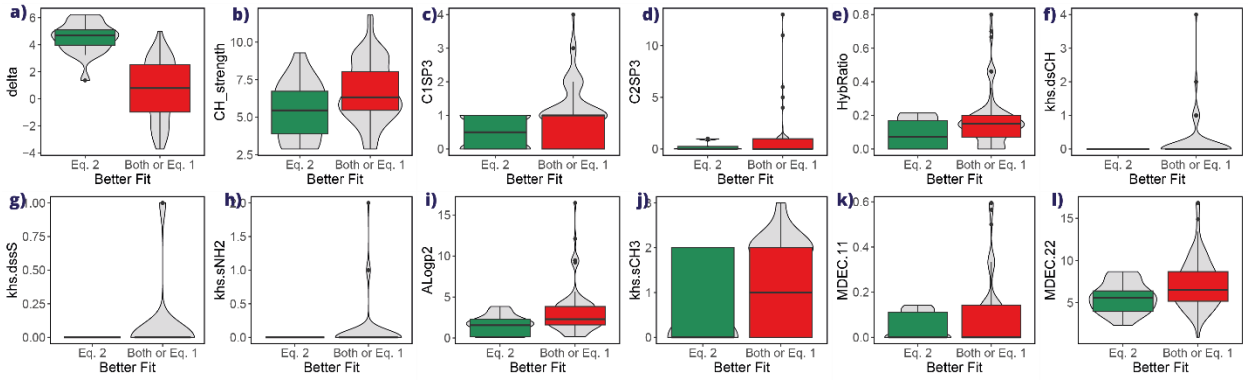


Figure 6. Violin plots of the distribution of the acidic molecules in our dataset along the selected descriptors for the acids ((a) *delta*, (b) *CH_strength*, (c) *C1SP3*, (d) *C2SP3*, (e) *HybRatio*, (f) *khs.dsCH*, (g) *khs.dssS*, (h) *khs.sNH2*, (i) *Alogp2*, (j) *khs.sCH3*, (k) *MDEC.11*, and (l) *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).

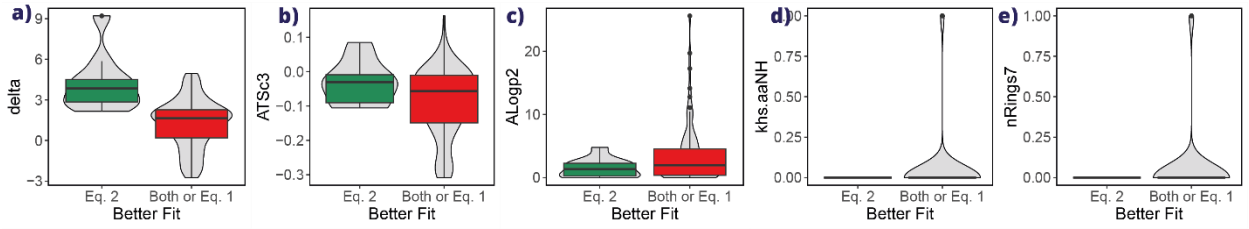


Figure 7. Violin plots of the distribution of the acidic molecules in our dataset along the selected descriptors for the bases (a) *delta*, (b) *ATSc3*, (c) *Alogp2*, (d) *khs.aaNH*, and (e) *nRings7*). Distributions are separated between acids and bases and classified by the binary operator $d_3 > 0.2$ (green) and $d_3 < 0.2$ (red).

Figure 6 and Figure 7 show the selected descriptors for acids and bases used to train our classification ML models, respectively. These descriptors showed a statistically significant divergence between the means of both populations among 180 descriptors tested for acids and bases.

Both, acidic and basic compounds show significant differences in their means ($p < 0.005$ in WTT test) for the descriptors delta and Alogp2 (see Table 1). The descriptor *delta* was calculated at the respective pH of each entry for the acids and bases. As expected, this descriptor correlates with the prominence of ionic species in both phases. Therefore, the apparent ionic partition became more significant for entries with higher *delta* values (see Figures 6a and Figure 7a). This result is very promising, because despite being an experimental descriptor, there are computational methods to determine the pK_a that include first principles models[116-119] as well as machine learning tools[120,121], so 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 Fig. S9)

The *ALogp2* descriptor consists of a 3D-QSAR model by Ghose & Crippen (1986) that predicts a square value of clogP_N value by analyzing the presence of 110 structural fragments within the molecules.[104] Figure 6i and 7c show that molecules with hydrophobicity close to logP = 0 (with lower *Alogp2* values) tend to fit best with log$D_{Eq.2}$. Water and *n*-octanol are not miscible, yet a small amount of water can dissolve in octanol at room temperature (~ 2.9 mol/kg).[105] These hydrophilic molecules might be dragged by the dissolved water to the octanol phase along with ionic species; thus, the apparent ionic partition would have a higher importance in these molecules.

This affinity for water, at least for acidic compounds, was further demonstrated by the *CH_strength* descriptor (Figure 6b). This descriptor, calculated by *Jazzy*, predicts the hydrogen-bond donor strength in carbon atoms.[98] 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 6 present

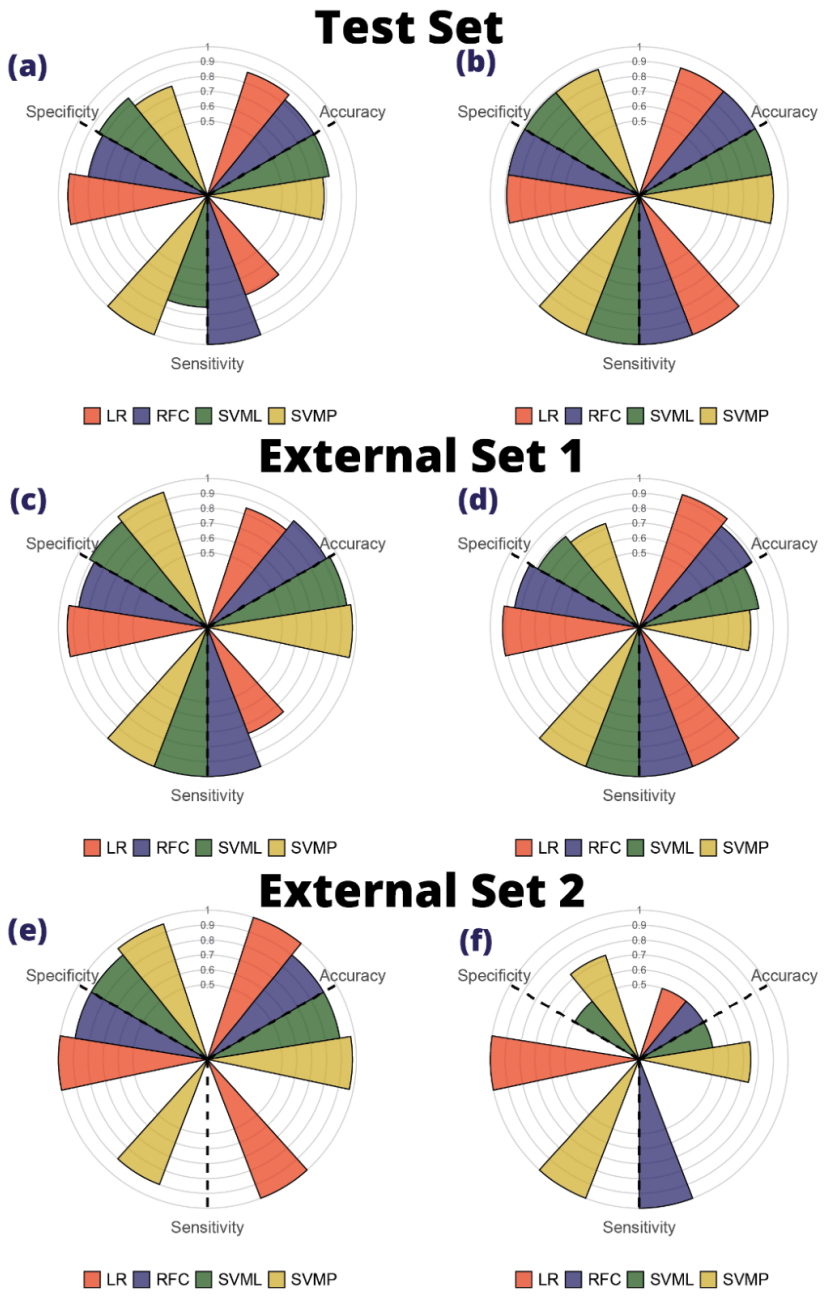
other important descriptor for acidic compounds such as *MDEC.22 and HypRatio.* The *MDEC.22* descriptor consists of a relationship between the number of secondary carbons in the molecule (i.e., vertices in a graph with only two paths) and the squared average atomic distance between those atoms.[101], whereas *HypRatio* 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 considering together the values of Alogp2, allows us to intuit that small and rigid ionizable molecules with instaurations or aromatic systems need considering the P_I^{app} to obtain an accurate prediction of $\log D_{pH.}$

Similarly, for basic compounds, higher values of ATSc3 descriptor are associated with taking into account the P_I^{app} for modeling pH-dependent lipophilic profiles. This descriptor is related with high molecular polarizability which agrees with the pattern of small molecules with the presence of polar atoms such as nitrogen.

Therefore, the apparent ionic partition effect should be considered for these small, rigid, and unsaturated molecules which present a significant proportion of ionic species in the aqueous phase species. It has been previously shown that the P_I^{app} of molecules may mechanistically occur via a simple ion-transfer reaction.[122] Thus, it is more plausible that small and compact molecules have a more prominent P_I^{app} because of the lower energetic cost of transferring to the cavity of the ion they replace.

Machine Learning Classification Models

Models A and B (see Figure 2) were trained using the LR, RFC, and SVM algorithms. A training set for acidic and basic molecules was used for each model and evaluated using the test set consisting of 20% of our population (see Figure S10). In addition, two external sets were validated with the experimental data of Disdier et al. (external set 1) [45] and Fauchère and Pliška (external set 2) [110]. Predictions were made as to 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 8 shows the results of the calculations for the four algorithms for the test and external sets of acidic and basic molecules.



534
535

536 **Figure 8.** Accuracy, sensitivity, and specificity of every ML model evaluated in this study for
537 acidic (a,c,e) and basic (b,d,f) entries within the test and external sets by defining our populations
538 with the conditional $d_3 > 0.2$. Descriptors were selected with the WTT method. Accuracies,
539 sensitivities, and specificities were calculated with Eqs. 9-11 based on the results of each confusion
540 matrix (Tables S1-S8)

It is observed that most of the calculated accuracies 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, it was observed that in the test set of acidic molecules, the sensitivity decreased, indicating that the models had difficulties in detecting molecules that fit log$D_{Eq.2}$ (Figure 7a). The external set related with capping amino acids reported by Fauchère and Pliška[110] 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 amount of phenolic groups. On the other hand, in the case of N-Acetyl-L-histidine amide, the results were very poor, this is due, at least in part, to the fact that our set has few bases in relation to the acids that best-fit to Eq. 2, and mainly because there was no imidazole fragment present in our set of bases, thus limiting the performance of our models.

Conclusions

Lipophilicity is undoubtedly the most used and important descriptor in the early stages of drug discovery and development. Additionally, it is a crucial descriptor in substance risk assessment and also in areas including adsorption in materials, catalysts, food chemistry, and computational biology. There are multiple tools to determine this descriptor, mainly for neutral molecules (logP_N), and for substances with ionizable groups, two formalisms are commonly used to determine the distribution coefficient (logD_{pH}), being the simplest pH correction model 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_I^{app}), since it has seen 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, validates the observations presented in limited previous studies. Thus, we develop machine learning algorithms using logistic regressions, random forest classifications, and support vector machine models to determine from molecular structures in which cases the 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, are better modeled using the formalism which takes into account the apparent ionic compounds (P_I^{app}).

Although we are aware of the molecular complexity of the species that can be included for the computational determination of the apparent ionic partition (P_I^{app}), parameterization or training of models using experimental values of P_I^{app} can help to 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 ionic partition (P_I^{app}) on the pH-dependent lipophilic profiles in more complex systems such as zwitterionic and peptides.

References

(1) Waring, M. J. Lipophilicity in Drug Discovery. *Expert Opin Drug Discov* **2010**, *5* (3), 235–248. https://doi.org/10.1517/17460441003605098.

(2) Lewis, D. F. V.; Jacobs, M. N.; Dickins, M. Compound Lipophilicity for Substrate Binding to Human P450s in Drug Metabolism. *Drug Discov Today* **2004**, *9* (12), 530–537. https://doi.org/https://doi.org/10.1016/S1359-6446(04)03115-0.

(3) Chatzopoulou, M.; Emer, E.; Lecci, C.; Rowley, J. A.; Casagrande, A. S.; Moir, L.; Squire, S. E.; Davies, S. G.; Harriman, S.; Wynne, G. M.; Wilson, F. X.; Davies, K. E.; Russell, A. J. Decreasing HepG2 Cytotoxicity by Lowering the Lipophilicity of Benzo[d]Oxazolephosphinate Ester Utrophin Modulators. *ACS Med Chem Lett* **2020**, *11* (12), 2421–2427. https://doi.org/10.1021/ACSMEDCHEMLETT.0C00405/ASSET/IMAGES/LARGE/ML0C00405_0003.JPEG.

(4) Miller, M. M.; Wasik, S. P.; Huang, G. L.; Shlu, W. Y.; Mackay, D. Relationships between Octanol-Water Partition Coefficient and Aqueous Solubility. *Environ Sci Technol* **1985**, *19* (6), 522–529. https://doi.org/10.1021/ES00136A007/ASSET/ES00136A007.FP.PNG_V03.

(5) Soliman, K.; Grimm, F.; Wurm, C. A.; Egner, A. Predicting the Membrane Permeability of Organic Fluorescent Probes by the Deep Neural Network Based Lipophilicity Descriptor DeepFl-LogP. *Scientific Reports 2021 11:1* **2021**, *11* (1), 1–9. https://doi.org/10.1038/S41598-021-86460-3.

(6) Esser, H. O. A Review of the Correlation between Physicochemical Properties and Bioaccumulation. *Pestic Sci* **1986**, *17* (3), 265–276. https://doi.org/10.1002/PS.2780170310.

(7) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and

Development Settings. *Adv Drug Deliv Rev* **2001**, *46* (1–3), 3–26. https://doi.org/10.1016/S0169-409X(00)00129-0.

(8) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J Med Chem* **2002**, *45* (12), 2615–2623. https://doi.org/10.1021/JM020017N/SUPPL_FILE/JM020017N_S.PDF.

(9) Leo, A.; Hansch, C.; Church, C. Comparison of Parameters Currently Used in the Study of Structure-Activity Relationships. *J Med Chem* **1969**, *12* (5), 766–771. https://doi.org/10.1021/JM00305A010/ASSET/JM00305A010.FP.PNG_V03.

(10) Leo, A.; Hansch, C.; Elkins, D. Partition Coefficients and Their Uses. *Chem Rev* **1971**, *71* (6), 525–616. https://doi.org/10.1021/CR60274A001/ASSET/CR60274A001.FP.PNG_V03.

(11) Kerns, E. H. High Throughput Physicochemical Profiling for Drug Discovery. *J Pharm Sci* **2001**, *90* (11), 1838–1858. https://doi.org/10.1002/JPS.1134.

(12) Liu, X.; Tu, M.; Kelly, R. S.; Chen, C.; Smith, B. J. Development of a Computational Approach to Predict Blood-Brain Barrier Permeability. *Drug Metab Dispos* **2004**, *32* (1), 132–139. https://doi.org/10.1124/DMD.32.1.132.

(13) Colmenarejo, G. In Silico Prediction of Drug-Binding Strengths to Human Serum Albumin. *Med Res Rev* **2003**, *23* (3), 275–301. https://doi.org/10.1002/MED.10039.(14)

(14) Zamora, W. J.; Curutchet, C.; Campanera, J. M.; Luque, F. J. Prediction of pH-Dependent Hydrophobic Profiles of Small Molecules from Miertus-Scrocco-Tomasi Continuum Solvation Calculations. *Journal of Physical Chemistry B* **2017**, *121* (42), 9868–9880.

(15) Iglesias, V., Pintado-Grima, C., Santos, J., Fornt, M., Ventura, S. Prediction of the Effect of pH on the Aggregation and Conditional Folding of Intrinsically Disordered Proteins with SolupHred and DispHred. In: Carugo, O., Eisenhaber, F. (eds) Data Mining Techniques for the Life Sciences. Methods in Molecular Biology. **2022**. vol 2449. Humana, New York, NY. https://doi.org/10.1007/978-1-0716-2095-3_8.

(16) Oeller M, Kang R, Bell R, Ausserwöger H, Sormanni P, Vendruscolo M. Sequence-based prediction of pH-dependent protein solubility using CamSol. Brief Bioinform. **2023**, 19;24(2):bbad004. doi: 10.1093/bib/bbad004.

(17) Porto WF, Ferreira KCV, Ribeiro SM, Franco OL. Sense the moment: A highly sensitive antimicrobial activity predictor based on hydrophobic moment. Biochim Biophys Acta Gen Subj. 2022 Mar;1866(3):130070. doi: 10.1016/j.bbagen.2021.130070.

(18) Zamora, W. J.; Pinheiro, S.; Separovic, F.; Luque, F. J. Insights into the Effect of the Membrane Environment on the Three-dimensional Structure-function Relationship of Antimicrobial Peptides. Biophysical Journal, 2020, 118, 3, 236a

(19) Simm, S.; Einloft, J.; Mirus, O.; Schleiff, E. 50 Years of Amino Acid Hydrophobicity Scales: Revisiting the Capacity for Peptide Classification. *Biol Res* **2016**, *49* (1), 1–19. https://doi.org/10.1186/S40659-016-0092-5.

(20) Zamora, W. J.; Campanera, J. M.; Luque, F. J. Development of a Structure-Based, PH-Dependent Lipophilicity Scale of Amino Acids from Continuum Solvation Calculations. *Journal of Physical Chemistry Letters* **2019**, *10* (4), 883–889.

(21) Ingram, T., Richter, U., Mehling, T., & Smirnova, I. Modelling of pH-dependent n-octanol/water partition coefficients of ionizable pharmaceuticals. Fluid Phase Equilibria. 2011, 305, 197-203.

(22) Chen, Chao and Shiang-Tai Lin. "Prediction of pH Effect on the Octanol–Water Partition Coefficient of Ionizable Pharmaceuticals." *Industrial & Engineering Chemistry Research. 2016,* 55, 9284-9294.

(23) Westall, J. C.; Leuenberger, C.; Schwarzenbach, R. P. Influence of pH and Ionic Strength on the Aqueous-Nonaqueous Distribution of Chlorinated Phenols. *Environ Sci Technol* **1985**, *19* (2), 193–198.

(24) Xing, L.; Glen, R. C. Novel Methods for the Prediction of LogP, pKa, and LogD. *J Chem Inf Comput Sci* **2002**, *42* (4), 796–805.

(25) Tetko, I. V.; Poda, G. I. Application of ALOGPS 2.1 to Predict Log D Distribution Coefficient for Pfizer Proprietary Compounds. *J Med Chem* **2004**, *47* (23), 5601–5604.

(26) Livingstone, D. Theoretical Property Predictions. *Curr Top Med Chem* **2005**, *3* (10), 1171–1192. https://doi.org/10.2174/1568026033452078

(27) C. C. Bannan, G. Calabro, D. Y. Kyu and D. L. Mobley, *J. Chem. Theory Comput*., 2016, **12**, 4015–4024.

(28) Bergazin, T. D.; Tielker, N.; Zhang, Y.; Mao, J.; Gunner, M. R.; Francisco, K.; Ballatore, C.; Kast, S. M.; Mobley, D. L. Evaluation of Log P, PK a, and Log D Predictions from the SAMPL7 Blind Challenge. *J Comput Aided Mol Des* **2021**, *35* (7), 771–802. https://doi.org/10.1007/s10822-021-00397-3.

(29) Avdeef, A. *Absorption and Drug Development.*; John Wiley & Sons, **2012.**

(30) Port, A.; Bordas, M.; Enrech, R.; Pascual, R.; Rosés, M.; Ràfols, C.; Subirats, X.; Bosch, E. Critical Comparison of Shake-Flask, Potentiometric and Chromatographic Methods for Lipophilicity Evaluation (Log P O/W) of Neutral, Acidic, Basic, Amphoteric, and Zwitterionic Drugs. *European Journal of Pharmaceutical Sciences* **2018**, *122*, 331–340. DOI:10.1016/j.ejps.2018.07.010.

(31) Austin, Rupert P. et al. The effect of ionic strength on liposome-buffer and 1-octanol-buffer distribution coefficients. Journal of Pharmaceutical Sciences. **1998**, 87, 5, 599-607 .

(32) Jain P, Kumar A. Concentration-dependent apparent partition coefficients of ionic liquids possessing ethyl- and bi-sulphate anions. Phys Chem Chem Phys. **2016**, 14;18(2):1105-13. doi: 10.1039/c5cp06611e.

(33) Jafvert, C. T., Westall, J. C., Grieder, E. & Schwarzenbach, R. P. Distribution of Hydrophobic Ionogenic Organic Compounds Between Octanol and Water: Organic Acids. *Environ. Sci. Technol.* **1990**, **24,** 1795–1803.

(34) Austin, R. P., Davis, a. M. & Manners, C. N. Partitioning of Ionizing Molecules between Aqueous Buffers and Phospholipid Vesicles. *J. Pharm. Sci.* **1995, 84,** 1180–1183.

(35) Takács-Novák, K. & Szász, G. Ion-Pair Partition of Quaternary Ammonium Drugs: The Influence of Counter Ions of Different Lipophilicity, Size, and Flexibility. *Pharm. Res.* **1999, 16,** 1633–1638.

(36) Fini, A. et al. Formation of Ion-Pairs in Aqueous Solutions of Diclofenac Salts. *Int. J. Pharm.* **1999**, **187,** 163–173.

(37) Sarveiya, V., Templeton, J. F. & Benson, H. a E. Ion-Pairs of Ibuprofen: Increased Membrane Diffusion. *J. Pharm. Pharmacol.* **2004, 56,** 717–724.

(38) Scherrer, R. A. & Donovan, S. F. Automated Potentiometric Titrations in KCl/Water-Saturated Octanol: Method for Quantifying Factors Influencing Ion-Pair Partitioning. *Anal. Chem.* **2009**, **81,** 2768–2778

(39) Wenlock, M. C., Potter, T., Barton, P. & Austin, R. P. A Method for Measuring the Lipophilicity of Compounds in Mixtures of 10. *J. Biomol. Screen*. **2011,** 16, 348–55.

(40) Fini, A., Bassini, G., Monastero, A. & Cavallari, C. Diclofenac Salts, VIII. Effect of the Counterions on the Permeation through Porcine Membrane from Aqueous Saturated Solutions. *Pharmaceutics*. **2012**, **4,** 413–429

(41) Paternostre M, Meyer O, Grabielle-Madelmont C, Lesieur S, Ghanam M, Ollivon M. Partition coefficient of a surfactant between aggregates and solution: application to the micelle-vesicle transition of egg phosphatidylcholine and octyl beta-D-glucopyranoside. *Biophys J.* 1995 Dec;69(6):2476-88. doi: 10.1016/S0006-3495(95)80118-9.

(42) Pieńko T, Grudzień M, Taciak PP, Mazurek AP. Cytisine basicity, solvation, logP, and logD theoretical determination as tool for bioavailability prediction. J Mol Graph Model. **2016** ;63:15-21. doi: 10.1016/j.jmgm.2015.11.003.

(43) Hung, L. Q. Electrochemical Properties of the Interface between Two Immiscible Electrolyte Solutions. *J. Electroanal. Chem.* **1980**, 115, 159–174.

(44) Kakiuchi, T. Limiting Behavior in Equilibrium Partitioning of Ionic Components in Liquid−Liquid Two-Phase Systems. *Anal. Chem.* **1996**, 68, 3658.

(45) Disdier, Z.; Savoye, S.; Dagnelie, R. V. H. Effect of Solutes Structure and PH on the N-Octanol/Water Partition Coefficient of Ionizable Organic Compounds. *Chemosphere* **2022**, *304*. https://doi.org/10.1016/j.chemosphere.2022.135155.

(46) Berthod, A.; Carda-Broch, S.; Garcia-Alvarez-Coque, M. C. Hydrophobicity of Ionizable Compounds. A Theoretical Study and Measurements of Diuretic Octanol-Water Partition Coefficients by Countercurrent Chromatography. *Anal Chem* **1999**, *71* (4), 879–888.

(47) Âde, F.; Reymond, R.; Carrupt, P.-A.; Testa, B.; Girault, H. H. Charge and Delocalisation Effects on the Lipophilicity of Protonable Drugs. *Chemistry – A European Journal* **1999**, *5* (1), 39–48.

(48) Gobry, V.; Ulmeanu, S.; Reymond, F.; Bouchard, G.; Carrupt, P. A.; Testa, B.; Girault, H. H. Generalization of Ionic Partition Diagrams to Lipophilic Compounds and to Biphasic Systems with Variable Phase Volume Ratios. *J Am Chem Soc* **2001**, *123* (43), 10684–10690.

(49) Reymond, F.; Steyaert, G.; Carrupt, P. A.; Testa, B.; Girault, H. Ionic Partition Diagrams: A Potential-PH Representation. *J Am Chem Soc* **1996**, *118* (47), 11951–11957.

(50) Cunha, R. D.; Ferreira, L. J.; Orestes, E.; Coutinho-Neto, M. D.; de Almeida, J. M.; Carvalho, R. M.; Maciel, C. D.; Curutchet, C.; Homem-de-Mello, P. Naphthenic Acids Aggregation: The Role of Salinity. *Computation* **2022**, *10* (10), 170.

(51) Tshepelevitsh, S.; Hernits, K.; Leito, I. Prediction of Partition and Distribution Coefficients in Various Solvent Pairs with COSMO-RS. *J Comput Aided Mol Des* **2018**, *32* (6), 711–722.

(52) Sonia Losada-Barreiro, Fátima Paiva-Martins, Carlos Bravo-Díaz. Partitioning of Antioxidants in Edible Oil–Water Binary Systems and in Oil-in-Water Emulsions. Antioxidants **2023**, *12* (4) , 828. https://doi.org/10.3390/antiox12040828

(53) Escher BI, Abagyan R, Embry M, Klüver N, Redman AD, Zarfl C, Parkerton TF. Recommendations for Improving Methods and Models for Aquatic Hazard Assessment of Ionizable Organic Chemicals. Environ Toxicol Chem. 2020 Feb;39(2):269-286. doi: 10.1002/etc.4602.

(54) Hansima MACK, Zvomuya F, Amarakoon I. Fate of veterinary antimicrobials in Canadian prairie soils - A critical review. *Science of The Total Environment* **2023**, *892* , 164387. https://doi.org/10.1016/j.scitotenv.2023.164387

(55) Tsantili-Kakoulidou, A.; Panderi, I.; Csizmadia, F.; Darvas, F. Prediction of Distribution Coefficient from Structure. 2. Validation of Prolog D, an Expert System. *American Pharmaceutical Association* **1997**, *86* (10), 1173–1179.

(56) Zamora, W. J.; Bertsch, E.; Suñer, S.; Pinheiro, S. Experimental n-Octanol/Water Partition/Distribution Coefficients Database for Small Molecules. **2023**. https://doi.org/10.5281/ZENODO.7956685.

(57) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res* **2021**, *49* (D1), D1388–D1395. https://doi.org/10.1093/NAR/GKAA971

(58) Mishra, B.; Sankar, C.; Mishra, M. Polymer Based Solutions of Bupranolol Hydrochloride for Intranasal Systemic Delivery. *J Drug Target* **2011**, *19* (3), 204–211. https://doi.org/10.3109/1061186X.2010.492520.

(59) Bello, M. L.; Junior, A. M.; Freitas, C. A.; Moreira, M. L. A.; da Costa, J. P.; de Souza, M. A.; Santos, B. A. M. C.; de Sousa, V. P.; Castro, H. C.; Rodrigues, C. R.; Cabral, L. M. Development of Novel Montmorillonite-Based Sustained Release System for Oral Bromopride Delivery. *European Journal of Pharmaceutical Sciences* **2022**, *175*. https://doi.org/10.1016/j.ejps.2022.106222.

(60) Bezençon, J.; Wittwer, M. B.; Cutting, B.; Smieško, M.; Wagner, B.; Kansy, M.; Ernst, B. PKa Determination by 1H NMR Spectroscopy - An Old Methodology Revisited. *J Pharm Biomed Anal* **2014**, *93*, 147–155. https://doi.org/10.1016/j.jpba.2013.12.014.

(61) Voigt, W.; Mannhold, R.; Limberg, J.; Blaschke, G. Interactions of Antiarrhythmics with Artificial Phospholipid Membranes. *J Pharm Sci* **1988**, *77* (12), 1018–1020.

(62) Roseman', T. J.; Yalkowsky, S. H. Physicochemical Properties of Prostaglandin FZa (Tromethamine Salt): Solubility Behavior, Surface Properties, and Ionization Constants. *J Pharm Sci* **1973**, *62* (10), 1680–1685.

(63) Shalaeva, M.; Kenseth, J.; Lombardo, F.; Bastin, A. Measurement of Dissociation Constants (PKa Values) of Organic Compounds by Multiplexed Capillary Electrophoresis Using

765 Aqueous and Cosolvent Buffers. *J Pharm Sci* **2008**, *97* (7), 2581–2606.
766 https://doi.org/10.1002/jps.21287.

767 (64) Qiang, Z.; Adams, C. Potentiometric Determination of Acid Dissociation Constants (PK a)
768 for Human and Veterinary Antibiotics. *Water Res* **2004**, *38* (12), 2874–2890.
769 https://doi.org/10.1016/j.watres.2004.03.017.

770 (65) de Almeida Drumond dos Santos, T.; Oliveira da Costa, D.; Silva da Rocha Pita, S.; Silva
771 Semaan, F. Potentiometric and Conductimetric Studies of Chemical Equilibria for Pyridoxine
772 Hydrochloride in Aqueous Solutions: Simple Experimental Determination of PKa Values and
773 Analytical Applications to Pharmaceutical Analysis. *Ecl. Quím* **2010**, *35* (4), 81–86.

774 (66) Morimotol, K.; Nagayasu, A.; Fukanoki, S.; Morisaka, K.; Hyon2, S.-H.; Ikada, Y.
775 Evaluation of Polyvinyl Alcohol Hydrogel as Sustained-Release Vehicle for Transdermal System
776 of Bunitrolol-HC1-1 1. *Drug Dev Ind Pharm* **1990**, *16* (1), 13–29.

777 (67) Loftsson, T.; Thorisdóttir, S.; Fridriksdóttir, H.; Stefánsson, E. Enalaprilat and Enalapril
778 Maleate Eyedrops Lower Intraocular Pressure in Rabbits. *Acta Ophthalmol* **2010**, *88* (3), 337–341.
779 https://doi.org/10.1111/j.1755-3768.2008.01495.x.

780 (68) Mannhold, R.; Dross, K. P.; FRekker, R.; der Steen, van. Drug Lipophilicity in QSAR
781 Practice: I. A Comparison of Experimental with Calculative Approaches. *Quant. Struct. Act. Relat.*
782 **1990**, *9*, 21–28.

783 (69) Loftsson, T.; Vogensen, S. B.; Desbos, C.; Jansook, P. Carvedilol: Solubilization and
784 Cyclodextrin Complexation: A Technical Note. *AAPS PharmSciTech* **2008**, *9* (2), 425–430.
785 https://doi.org/10.1208/s12249-008-9055-7.

786 (70) Kuntworbe, N.; Alany, R. G.; Brimble, M.; Al-Kassas, R. Determination of PKa and
787 Forced Degradation of the Indoloquinoline Antimalarial Compound Cryptolepine Hydrochloride.
788 *Pharm Dev Technol* **2013**, *18* (4), 866–876. https://doi.org/10.3109/10837450.2012.668554.

789 (71) Islam, M. S.; Narurkar, M. M. Solubility, Stability and Ionization Behaviour of Famotidine.
790 *Journal of Pharmacy and Pharmacology* **1993**, *45* (8), 682–686. https://doi.org/10.1111/j.2042-
791 7158.1993.tb07088.x.

792 (72) Deng, Y.; Li, B.; Yu, K.; Zhang, T. Biotransformation and Adsorption of Pharmaceutical
793 and Personal Care Products by Activated Sludge after Correcting Matrix Effects. *Science of the*
794 *Total Environment* **2016**, *544*, 980–986. https://doi.org/10.1016/j.scitotenv.2015.12.010.

795 (73) Franke, U.; Munk, A.; Wiese, M. Ionization Constants and Distribution Coefficients of
796 Phenothiazines and Calcium Channel Antagonists Determined by a PH-Metric Method and
797 Correlation with Calculated Partition Coefficients. *J Pharm Sci* **1999**, *88* (1), 89–95.
798 https://doi.org/10.1021/js980206m.

799 (74) Avdeef, A.; Box, K. J.; Comer, J. E. A.; Hibbert, C.; Tam, K. Y. PH-Metric LogP 10.
800 Determination of Liposomal Membrane-Water Partition Coefficient of Ionizable Drugs. *Pharm*
801 *Res* **1998**, *15* (2), 209–215.

802 (75) Thanacoody, R. H. K. Thioridazine: The Good and the Bad. *Recent Pat Antiinfect Drug*
803 *Discov* **2011**, *6*, 92–98.

(76) Martínez, V.; Maguregui, M. I.; Jiménez, R. M.; Alonso, R. M. Determination of the PK a Values of B-Blockers by Automated Potentiometric Titrations. *J Pharm Biomed Anal* **2000**, *23*, 459–468.

(77) Huerta, B.; Jakimska, A.; Gros, M.; Rodríguez-Mozaz, S.; Barceló, D. Analysis of Multi-Class Pharmaceuticals in Fish Tissues by Ultra-High-Performance Liquid Chromatography Tandem Mass Spectrometry. *J Chromatogr A* **2013**, *1288*, 63–72. https://doi.org/10.1016/j.chroma.2013.03.001.

(78) Fini, A.; Fazio, G.; Feroci, G. Solubility and Solubilization Properties of Non-Steroidal Anti-Inflammatory Drugs. *Int J Pharm* **1995**, *126* (1–2), 95–102. https://doi.org/10.1016/0378-5173(95)04102-8.

(79) Jacka, M. R. *Clarke's Isolation and Identification of Drugs*, 2nd ed.; Moffat, A. C., Jackson, J. V., Moss, M. S., Widdop, B., Greenfield, E. S., Eds.; Pharmaceutical Press: London, 2000.

(80) Nakamura, Y.; Yamamoto, H.; Sekizawa, J.; Kondo, T.; Hirai, N.; Tatarazako, N. The Effects of PH on Fluoxetine in Japanese Medaka (Oryzias Latipes): Acute Toxicity in Fish Larvae and Bioaccumulation in Juvenile Fish. *Chemosphere* **2008**, *70* (5), 865–873. https://doi.org/10.1016/j.chemosphere.2007.06.089.

(81) Schräder, W.; Andersson, J. T. Fast and Direct Method for Measuring 1-Octanol-Water Partition Coefficients Exemplified for Six Local Anesthetics. *J Pharm Sci* **2001**, *90* (12), 1948–1954. https://doi.org/10.1002/JPS.1145.

(82) Avdeef, A. *Sirius Technical Application Notes (STAN)*; Sirius Analytical Instruments Ltd.: Forest Row, UK, 1994; Vol. 1.

(83) Avdeef, A.; Box, K. J.; Comer, J. E. A.; Hibbert, C.; Tam, K. Y. PH-Metric LogP 10. Determination of Liposomal Membrane-Water Partition Coefficients of Ionizable Drugs. *Pharm Res* **1998**, *15* (2), 209–215. https://doi.org/10.1023/A:1011954332221/METRICS.

(84) Caron, G.; Steyaert, G.; Pagliara, A.; Âde, F.; Reymond, Â.; Crivori, P.; Gaillard, P.; Carrupt, P.-A.; Avdeef, A.; Comer, J.; Box, K. J.; Girault, H. H.; Testa, B. Structure-Lipophilicity Relationships of Neutral and Protonated b-Blockers Intra-and Intermolecular Effects in Isotropic Solvent Systems. https://doi.org/10.1002/(SICI)1522-2675(19990804)82:8.

(85) Avdeef, A. *Sirius Technical Application Notes (STAN)*; Sirius Analytical Instruments Ltd.: Forest Row, UK, 1995; Vol. 2.

(86) Lombardo, F.; Shalaeva, M. Y.; Tupper, K. A.; Gao, F.; Abraham, M. H. ElogP(Oct): A Tool for Lipophilicity Determination in Drug Discovery. *J Med Chem* **2000**, *43* (15), 2922–2928. https://doi.org/10.1021/JM0000822/ASSET/IMAGES/MEDIUM/JM0000822E00013.GIF.

(87) Winiwarter, S.; Bonham, N. M.; Ax, F.; Hallberg, A.; Lennernäs, H.; Karlén, A. Correlation of Human Jejunal Permeability (in Vivo) of Drugs with Experimentally and Theoretically Derived Parameters. A Multivariate Data Analysis Approach. *J Med Chem* **1998**, *41* (25), 4939–4949. https://doi.org/10.1021/JM9810102/SUPPL_FILE/JM9810102_S.PDF.

(88) Slater, B.; McCormack, A.; Avdeef, A.; Comer, J. E. A. PH-Metric Log P. 4. Comparison of Partition Coefficients Determined by HPLC and Potentiometric Methods to Literature Values. *J Pharm Sci* **1994**, *83* (9), 1280–1283. https://doi.org/10.1002/JPS.2600830918.

(89) Luger, P.; Daneck, K.; Engel, W.; Trummlitz, G.; Wagner, K. Structure and Physicochemical Properties of Meloxicam, a New NSAID. *European Journal of Pharmaceutical Sciences* **1996**, *4* (3), 175–187. https://doi.org/10.1016/0928-0987(95)00046-1.

(90) Takács-Novák, K.; Józan, M.; Hermecz, I.; Szász, G. Lipophilicity of Antibacterial Fluoroquinolones. *Int J Pharm* **1992**, *79* (1–3), 89–96. https://doi.org/10.1016/0378-5173(92)90099-N.

(91) Carda-Broch, S.; Berthod, A. PH Dependence of the Hydrophobicity of β-Blocker Amine Compounds Measured by Counter-Current Chromatography. *J Chromatogr A* **2003**, *995* (1–2), 55–66. https://doi.org/10.1016/S0021-9673(03)00534-X.

(92) Scott, D. Estimation of Distribution Coefficients from the Partition Coefficient and PKa. *Pharmaceutical Technology* **2002**, *26* (11).

(93) Gulaboski, R.; Borges, F.; Pereira, C. M.; Natália, M.; Cordeiro, D. S.; Garrido, J.; Silva, A. F. Voltammetric Insights in the Transfer of Ionizable Drugs Across Biomimetic Membranes-Recent Achievements. *Comb Chem High Throughput Screen* **2007**, *10*, 514–526.

(94) *pKa Plugin | Chemaxon Docs*. https://docs.chemaxon.com/display/docs/pka-plugin.md (accessed 2023-01-24).

(95) *LogP and logD calculations | Chemaxon Docs*. https://docs.chemaxon.com/display/docs/logp-and-logd-calculations.md (accessed 2023-01-24).

(96) https://cran.r-project.org/web/packages/Metrics/Metrics.pdf

(97) E. L. Willighagen, J. W. Mayfield, J. Alvarsson, A. Berg, L. Carlsson, N. Jeliazkova, S. Kuhn, T. Pluskal, M. Rojas-Chertó, O. Spjuth, G. Torrance, C. T. Evelol, R. Guha and G. Steinbeck, *J. Cheminform*., **2017**, **9**, 1–19.

(98) Ghiandoni, G. M.; Caldeweyher, E. Fast Calculation of Hydrogen-Bond Strengths and Free Energy of Hydration of Small Molecules. *Scientific Reports 2023 13:1* **2023**, *13* (1), 1–11. https://doi.org/10.1038/S41598-023-30089-X.

(99) Welch, B. L. The Generalization Of 'Student's' Problem When Several Different Population Variances Are Involved. *Biometrika* **1947**, *34* (1–2), 28–35. https://doi.org/10.1093/BIOMET/34.1-2.28.

(100) Ruxton, G. D. The Unequal Variance T-Test Is an Underused Alternative to Student's t-Test and the Mann–Whitney U Test. *Behavioral Ecology* **2006**, *17* (4), 688–690. https://doi.org/10.1093/BEHECO/ARK016.

(101) Kuhn, M.; Johnson, K. *Feature Engineering and Selection: A Practical Approach for Predictive Models*, 1st ed.; Taylor and Francis Group: London, 2019.

(102) https://cran.r-project.org/web/packages/caret/

(103) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance-Edge (MDE) Vector, γ. *J Chem Inf Comput Sci* **1998**, *38* (3), 387–394.

(104) Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a

884 Measure of Hydrophobicity. *J Comput Chem* **1986**, *7* (4), 565–577.
885 https://doi.org/10.1002/JCC.540070419.

886 (105) Šegatin, N.; Klofutar, C. Thermodynamics of the Solubility of Water in 1-Hexanol, 1-
887 Octanol, 1-Decanol, and Cyclohexanol. *Monatsh Chem* **2004**, *135* (3), 241–248.

888 (106) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel
889 Combination of Electronic, Topological, and Valence State Information. *J Chem Inf Comput Sci*
890 **1995**, *35* (6), 1039–1045.

891 (107) https://cran.r-project.org/web/packages/bestglm/bestglm.pdf

892 (108) Furnival, G. M.; Wilson, R. W. Regressions by Leaps and Bounds. *Technometrics* **1974**,
893 *16* (4), 499–511. https://doi.org/10.1080/00401706.1974.10489231.

894 (109) Burger, S. *Introduction to Machine Learning with R: Rigorous Mathematical Modeling*,
895 1st ed.; O'Reilly: United States of America, 2018.

896 (110) Fauchere, J. L. & Pliska, V. Hydrophobic Parameters π of Amino Acid Side Chains
897 from the Partitioning of *N*-Acetyl-*L*-Amino Acid Amides. *Eur. J. Med. Chem.* **1983, 18,** 369–
898 375.

899 (111) Breiman, L. Random Forests. *Mach Learn* **2001**, *45* (1), 5–32.
900 https://doi.org/10.1023/A:1010933404324/METRICS.

901 (112) https://cran.r-project.org/web/packages/randomForest/index.html.

902 (113) Cortes, C.; Vapnik, V.; Saitta, L. Support-Vector Networks. *Machine Learning 1995 20:3*
903 **1995**, *20* (3), 273–297. https://doi.org/10.1007/BF00994018.

904 (114) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. Training Algorithm for Optimal Margin
905 Classifiers. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*
906 **1992**, 144–152. https://doi.org/10.1145/130385.130401.

907 (115) https://cran.r-project.org/web/packages/e1071/index.html

908 (116) Viayna A, Antermite SG, de Candia M, Altomare CD, Luque FJ. Interplay between
909 Ionization and Tautomerism in Bioactive β-Enamino Ester-Containing Cyclic Compounds: Study
910 of Annulated 1,2,3,6-Tetrahydroazocine Derivatives. J Phys Chem B. **2020** Jan 9;124(1):28-37.
911 doi: 10.1021/acs.jpcb.9b08904.

912 (117) Tielker N, Güssregen S, Kast SM. SAMPL7 physical property prediction from EC-RISM
913 theory. J Comput Aided Mol Des. **2021** Aug;35(8):933-941. doi: 10.1007/s10822-021-00410-9.

914 (118) Viayna A, Pinheiro S, Curutchet C, Luque FJ, Zamora WJ. Prediction of n-octanol/water
915 partition coefficients and acidity constants (pK_a) in the SAMPL7 blind challenge with the
916 IEFPCM-MST model. J Comput Aided Mol Des. **2021** Jul;35(7):803-811.

917 (119) Rodriguez SA, Tran JV, Sabatino SJ, Paluch AS. Predicting octanol/water partition
918 coefficients and pKa for the SAMPL7 challenge using the SM12, SM8 and SMD solvation models.
919 J Comput Aided Mol Des. **2022** Sep;36(9):687-705. doi: 10.1007/s10822-022-00474-1.

920 (120) Wu J, Kang Y, Pan P, Hou T. Machine learning methods for pK_a prediction of small
921 molecules: Advances and challenges. Drug Discov Today. **2022** Dec;27(12):103372. doi:
922 10.1016/j.drudis.2022.103372.

(121) Johnston RC, Yao K, Kaplan Z, Chelliah M, Leswing K, Seekins S, Watts S, Calkins D, Chief Elk J, Jerome SV, Repasky MP, Shelley JC. Epik: pK_a and Protonation State Prediction through Machine Learning. J Chem Theory Comput. **2023** ,25;19(8):2380-2388. doi: 10.1021/acs.jctc.3c00044.

(122) Reymond, F.; Chopineaux-Courtois, V.; Steyaert, G.; Bouchard, G.; Carrupt, P. A.; Testa, B.; Girault, H. H. Ionic Partition Diagrams of Ionisable Drugs: PH-Lipophilicity Profiles, Transfer Mechanisms and Charge Effects on Solvation. *Journal of Electroanalytical Chemistry* **1999**, *462* (2), 235–250. https://doi.org/10.1016/S0022-0728(98)00418-5

(123) Burden, F. R. Molecular Identification Number for Substructure Searches. *J Chem Inf Comput Sci* **1989**, *29* (3), 225–227.

(124) Burden, F. R. A Chemically Intuitive Molecular Index Based on the Eigenvalues of a Modified Adjacency Matrix. *Quant. Stmct-Act. Relat* **1997**, *16*, 3–314.

(125) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J Chem Inf Comput Sci* **1999**, *39* (1), 28–35.

(126) Petitjean, M. Applications of the Radius-Diameter Diagram to the Classification of Topological and Geometrical Shapes of Chemical Compounds. *J Chem Inf Comput Sci* **1992**, *32* (4), 331–337.

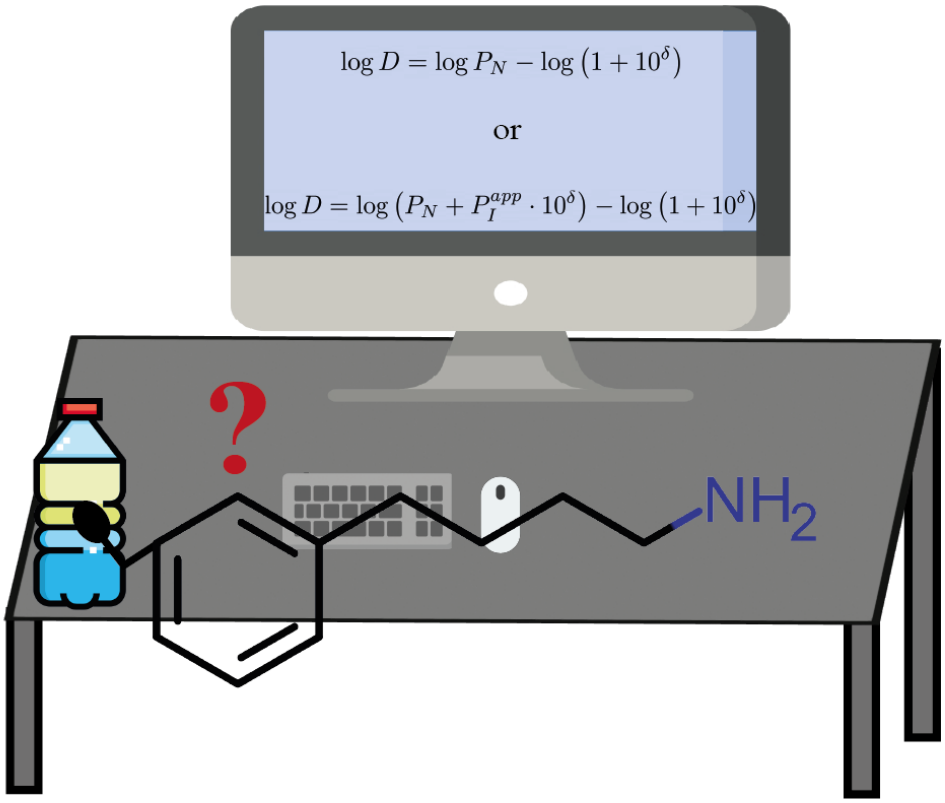
(127) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J Med Chem* **2000**, *43* (20), 3714–3717.

958

959 ***TOC Graphics***

960

961



962