LiProS: FAIR simulation workflow to Predict Accurate Lipophilicity Profiles for Small Molecules.

Esteban Bertsch-Aguilar[1,2], Antonio Piedra[1], Daniel Acuña[2], Sebastián Suñer[2], Sylvana Pinheiro[2,3], and William J. Zamora[1,2,3,*]

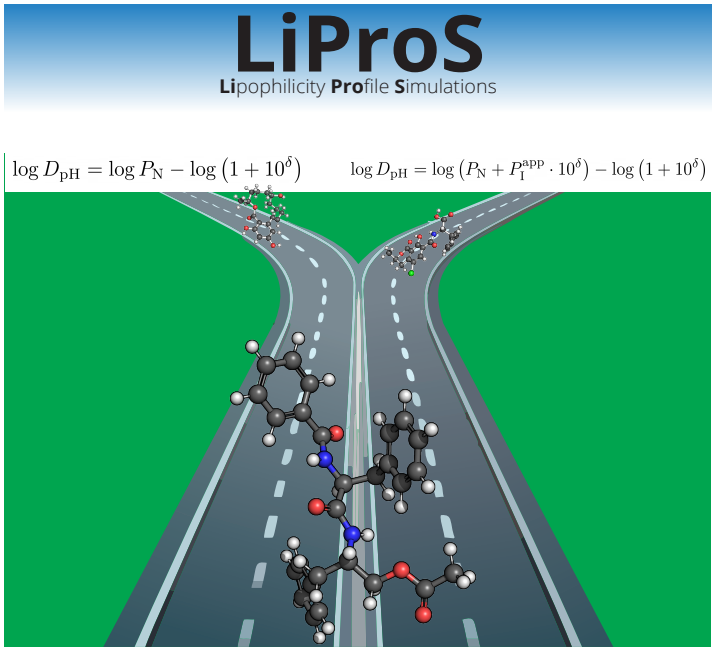
[1]National Advanced Computing Colaboratory (CNCA), National High Technology Center (CeNAT), Pavas, San José, Costa Rica.
[2]CBio[3] Laboratory, School of Chemistry, University of Costa Rica, San Pedro, San José, Consta Rica.
[3]Laboratory of Computational Toxicology and Artificial Intelligence (LaToxCIA), Biological Testing Laboratory (LEBi), University of Costa Rica, San Pedro, San José, Consta Rica.
[*]Corresponding Author: william.zamoraramirez@ucr.ac.cr
October 15, 2024

ABSTRACT

The consideration of the ionic partition coefficient in estimating pH-dependent lipophilicity profiles for small molecules has been previously emphasized through classification Machine Learning protocols. In alignment with the principles of *Findable, Accessible, Interoperable, and Reusable* (FAIR) data to enhance data management and sharing, we introduce LiProS: a FAIR workflow accessible via Google Colab. LiProS assists researchers in efficiently determining the appropriate pH-dependent lipophilicity profile based on the SMILES code of their molecules of interest. LiProS demonstrated its applicability in discerning the most suitable lipophilicity formalism based on small structural variations in potential cases of structure-based drug design.

TOC Abstract

LiProS is a FAIR-compliant workflow accessible via Google Colab, helps researchers quickly determine the appropriate profile based on the SMILES code of their molecules. LiProS has shown effectiveness in identifying the best lipophilicity formalism, and can be applied in structure-based drug design.



1 Introduction

Lipophilicity, the affinity of a molecule for a non-polar solvent relative to a polar solvent, is quantified by the partition coefficient (P_N), representing the equilibrium between an organic and aqueous phase.[1] While P_N can be measured in various biphasic systems, the *n*-octanol/water system is most widely used due to its relevance in drug design, environmental chemistry, and food chemistry.[2–8] Notably, $\log P_N$ correlates with key pharmacokinetic properties such as toxicity, membrane permeability, and bioaccumulation, and is central to Lipinski's *Rule of 5* for assessing oral bioavailability in drug candidates.[9–13]

To predict $\log P_N$, various *in silico* methods have been developed, including first-principles models and data-driven approaches like QSAR and Machine Learning (ML), though the former often involve high computational costs.[14–20] However, $\log P_N$ describes only the neutral species, whereas many drug-like molecules contain ionizable groups, making lipophilicity pH-dependent.[21] This pH-dependent lipophilicity is expressed as the distribution coefficient (D_{pH}), where ionized molecules typically show lower values than their neutral counterparts. Techniques like potentiometry and cyclic voltammetry have enabled the measurement of $\log D_{pH}$ across different pH values, which is crucial for understanding pharmacokinetic properties like brain cell permeability and protein binding.[22–29] Additionally, the lipophilic efficiency (LipE) of a molecule, which is calculated by subtracting the potency of a drug candidate (often its pIC^{50}) from its $\log D_{7.4}$, serves as a penalization factor to better assess bioavailability.[30] In food chemistry, the $\log D_{6.8}$ has been correlated with the distribution of drugs among various fractions (fat, curd, whey, etc.) in spiked milk samples, providing insight into the bioaccumulation potential of these drugs.[28] Similar applications have been identified in environmental chemistry, where the $\log D_{7.4}$ of molecules in PDMS/water systems has been measured, offering a better understanding of the passive equilibrium sampling of dissolved contaminants in water.[29]

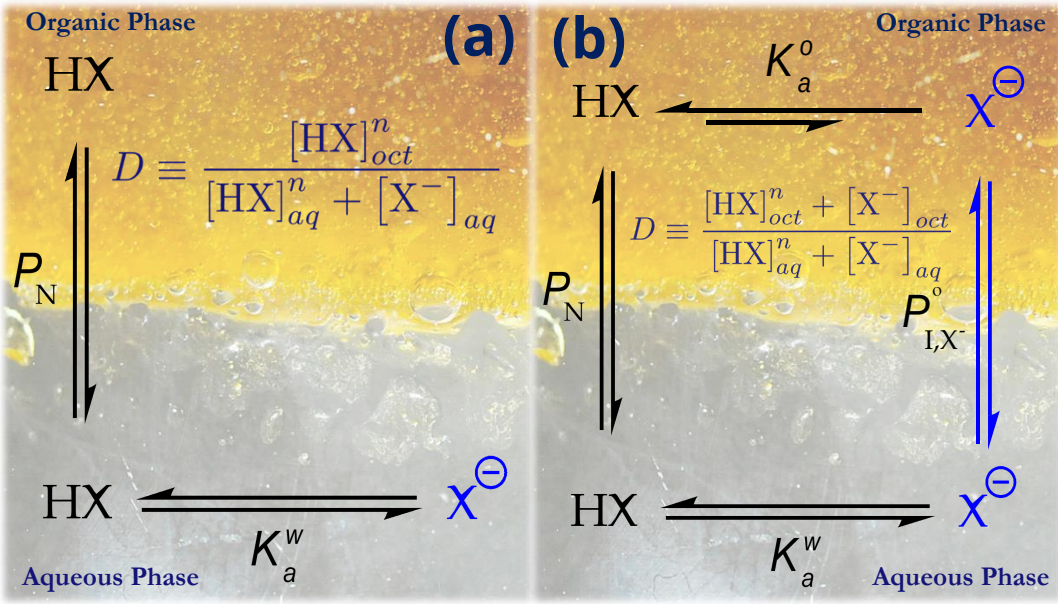


Figure 1. Representations of the partition mechanism for a symbolic acidic molecule its neutral (**HX**) and ionic (**X**$^-$) by (a) assuming the exclusive ionization in the aqueous phase, and (b) considering the partition of ions between phases.

This wide range of applications has led to the search for simple prediction models of lipophilicity profiles for small molecules. The use of thermodynamically derived equations for $\log D_{pH}$ has been successful in facilitating straightforward calculations. Fig. 1a represents the most commonly applied equilibrium model for the distribution coefficient of ionizable molecules. This model assumes exclusive ionization in the aqueous phase, simplifying the distribution coefficient as:

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

where δ = pH - pK_a for acids and δ = pK_a - pH for bases. This simple equation provides the $\log D_{pH}$ of a molecule using its $\log P_N$ and its pK_a. However, this equation does not accurately represent real-world environments and often diverges from experimental measurements.[31] Fig. 1b depicts a more accurate equilibrium model, where the ionic species partition between phases. The applied ionic partitioning (P_I^{app}) of species has been demonstrated in both theoretical and experimental studies.[32;33] This equilibrium simplifies to the following equation:

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

As shown above, lipophilicity is a cornerstone property for easily and efficiently assessing crucial parameters in drug design, materials and food engineering, and in scoring functions for modeling biomolecule interactions. As a result, there are many commercial software's available to determine this descriptor.[34] In line with the principles for *Findable, Accessible, Interoperable, and Reusable* data (FAIR) to improve data management and sharing, there is a trend to not only share small molecule lipophilicity data but also the workflows that allow FAIR principles to be applied to them.[35]

In previous work, we conducted a large-scale evaluation of Equations 1 and 2 using a dataset of small ionizable molecules.[31] Equation 2 generally produced more accurate results. However, for many molecules, the improvement was negligible, and thus the use of Equation 1 may be recommended in such cases. These findings prompted us to perform a critical assessment of both groups of molecules. We developed a simple ML classification algorithm to guide the determination, through molecular descriptors, of which equation will yield a more accurate $\log D_{\mathrm{pH}}$ calculation.

Herein, we further develop this ML model with an expanded database and a greater number of available descriptors. Additionally, we introduce **LiProS**: a FAIR workflow, easily accessible through Google Colab, with the aim of assisting researchers in quickly and easily determining which pH-dependent lipophilicity profile should be used for their molecules of interest.

2 Methodology

2.1 Classification Models Training

A previously curated dataset was used to train our models.[36] Experimental values of $\log P_{\mathrm{N}}$, $\log P_{\mathrm{I}}^{\mathrm{app}}$, $\mathrm{p}K_{\mathrm{a}}$, and $\log D_{\mathrm{pH}}$ at various pH levels were collected from the literature, along with the corresponding measurement techniques.

Molecular descriptors were calculated for each data entry's SMILES code using the `RCDK` library in R, as well as the `RDKit` and `Openbabel` packages in Python.[37–39] Additionally, we employed `Jazzy` for calculating solvation free energies and hydrogen-bond strengths, and `ConjugaR`, an in-house tool, for identifying conjugated systems.[40]

Acidic and basic molecules were separated, and key features were selected using a previously described method.[31] A total of 21 acidic and 34 basic descriptors were chosen to train our models (Table S1, Fig. S1). All models showed a statistically significant difference ($\alpha = 0.05$) between entries better fit by Eq. 1 versus Eq. 2 (Fig. S2-S3). The importance of the selected features was obtained through the `meandecreasGini` score (Fig. 3).

Datasets were randomly split into training (80%) and test (20%) sets. Logistic regression (LR), random forest (RF), and support vector machine (with a linear kernel, SVML) models were trained on the selected descriptors and evaluated on the test set. Model outputs were averaged to ensure balanced results, which were then stored in confusion matrices for the calculation of accuracy, sensitivity, and specificity. Furthermore, these models were compared to the best classification models developed by the ROBERT ML protocols tool, using descriptors calculated by the AQME package.[41;42]

All scripts, databases, and models were stored in a GitHub repository to ensure reproducibility (see Data availability).

2.2 FAIR Workflow via Google Colab

A Google Colab script was prepared to enhance the accessibility of our models. The workflow consists of the following steps:

1. **Tools Installation:** These steps only require the user's execution of each code cell, with no further input necessary.
   (a) Load all necessary packages to run the models.
   (b) Load the ML models from our repository. These models are imported from our GitHub repository (see Data availability). For our acidic model, the LR, RF, and SVML models were loaded with the aim of averaging their outputs to obtain the final result. For our basic model, only the SVML model was required (check Results and Discussion).
   (c) Execute the descriptor calculation function.
2. **Import Your Molecules:** The script is divided into sections for acidic and basic molecules. Therefore, users must separate their molecules beforehand. After executing the cell, the following input cells will appear below:
   (a) *Enter the SMILES of your ACIDIC/BASIC molecules:* Users must enter the SMILES code of the desired molecules, with each code separated by spaces (e.g., "`CC(=O)O O=C(O)c1ccccc1`" for acetic and benzoic acid, respectively). The SMILES code will be used to calculate the descriptors.

(b) *Enter the desired pH:* Users must enter the desired pH for each individual molecule, separated by spaces (e.g., to determine the best $\log D_{pH}$ formalism at physiological pH, users should enter "`7.4 7.4`").

(c) *Is the molecule acidic or basic?:* Users must declare if their molecules are acidic or basic by typing `acid` or `base`. If a molecule is ampholytic or zwitterionic, its acidic or basic behavior can be approximated using its pK_a values. In such cases, the user should review which ionic species predominates over a broader pH range to approximate the acidic or basic behavior.

(d) *Enter the molecule's pKa:* The user must enter each molecule's pK_a separated by spaces. If the experimental pK_a is unknown, one can estimate it using any predictive tool.

Otherwise, you can enter your data as an Excel (`.xlsx`) file. In this case, the user must import their database in the Google Colab environment. The file must have the following columns:

(a) **SMILES:** The SMILES code for each molecule.

(b) **type:** Is the molecule mostly acidic or basic?

(c) **pKa:** Insert the pK_a of the molecule. It can be an estimated value.

(d) **pH:** The desired pH of the molecule.

3. **Output:** These cells will calculate the descriptors for each molecule, and the ML models will predict the recommended $\log D_{pH}$ formalism ("`Use Eq.1`" or "`Use Eq.2`"). The information will be saved in a `.csv` file containing the SMILES code, pH, and output column. This file will be automatically downloaded to the user's computer.

2.3 Applications of LiProS

We employed LiProS in a drug design application. A natural progression was undertaken, beginning with the evaluation of the biocompatibility of three alkaloids. These ligands were assessed for their affinity scores with the proteins *Trypanothione reductase* and *Cruzain*, which are present in the parasite *Trypanosoma cruzi*, the causative agent of Chagas disease.[43] Google Colab was utilized as a proof-of-concept tool to demonstrate its utility in rapidly determining the most appropriate $\log D_{pH}$ formalism for each substance. Fig. 2 presents the selected molecules.

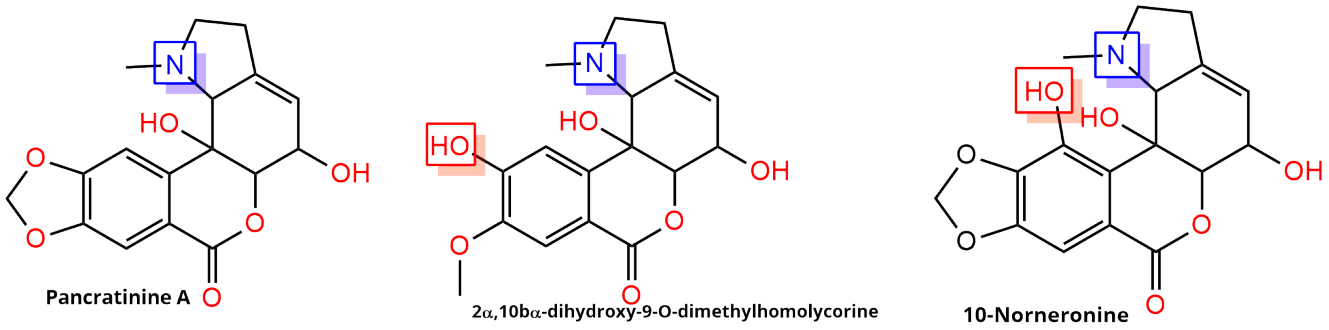


Figure 2. Structures of the alkaloids that had the highest docking score for the proposed binding sites of the *Trypanothione reductase* and *Cruzain* proteins.[43] Acidic and basic ionizable groups are labeled with red and blue squares, respectively.

Subsequently, the ability of LiProS to evaluate large datasets was also tested. For this purpose, the Natural Products Repository of Costa Rica (NAPRORE-CR) was used. This database is part of the Latin American Natural Product Database.[44;45] It contains approximately 932 natural products extracted from various organisms in Costa Rica. Each molecule is classified according to its pathways, classes, and the species from which it was obtained. Additionally, the SMILES code for each product is stored. Using this information, the database was scanned to select only the molecules with ionizable groups. For this, the tools implemented within the pIChemist package, developed by Frolov and collaborators, were employed.[46] The implementation of this tool to identify ionizable groups was added to our GitHub repository to ensure reproducibility (`ionizable-group-counter.ipynb`). As a result, 323 ionizable molecules were obtained. These molecules were then input into LiProS as a `.xlsx` file to predict, at a pH of 7.4, which lipophilicity formalism is most appropriate for calculating the $\log D_{7.4}$.

3 Results and Discussion

3.1 Classification models performance

The ML models utilized in this study function as simple classification algorithms to determine whether a molecule is more accurately described by Eq. 1 or Eq. 2. Topological, representative, and experimental descriptors were selected based on their ability to distinguish significantly between the two groups, as evaluated by their descriptor values (Table S1). As a result, univariate feature selection, relying on descriptor performance in a two-tailed *t*-test, was identified as the most suitable feature selection method (Figs. S2-S3). Although alternative iterative feature selection methods, such as recursive feature elimination, were explored in previous studies with satisfactory outcomes, these methods frequently resulted in overfitting.[31] Furthermore, the importance of each descriptor was assessed, and Fig. 3 provides a summary of the importance score for each feature. The experimental descriptor `delta` ($\delta = \text{pH} - \text{p}K_\text{a}$ or $\delta = \text{p}K_\text{a} - \text{pH}$) exhibited the highest importance, given the context of pH-dependent lipophilicity profiles. Higher δ values indicate a greater prevalence of ionic species in both aqueous and organic phases. The models predominantly base their decisions on the δ value, determined by the pK_a of the molecule and the desired pH. However, other descriptors also demonstrated significant relevance through various structural features. For instance, descriptors such as `rotors`, `MDEC.22`, `MDEO.11`, and several fragment-based topological descriptors exhibited expected behavior, indicating that smaller, more rigid molecules are more likely to align with Eq. 2. High hydrophilicity was also reflected in the consideration of P_I^app by our models (`logP Obabel` and `SlogP VSA5`). Moreover, descriptors related to polarizability, partial charge, and heteroatoms also influenced the model's decisions (`fpDensityMorgan 3`, `Estate VSA3`, `tpsaEfficiency`, etc.). In these instances, a higher presence of potentially ionizable heteroatoms increases the likelihood of the models selecting Eq. 2. These findings are consistent with our previous model, which concluded that *small, rigid, and unsaturated molecules with a* $\log P_N$ *close to zero, representing a significant proportion of ionic species in the aqueous phase, are better modeled using the formalism that accounts for the apparent ionic compounds* P_I^{app}.[31] Further details on these descriptors can be found in the documentation for each employed package.[37–40]

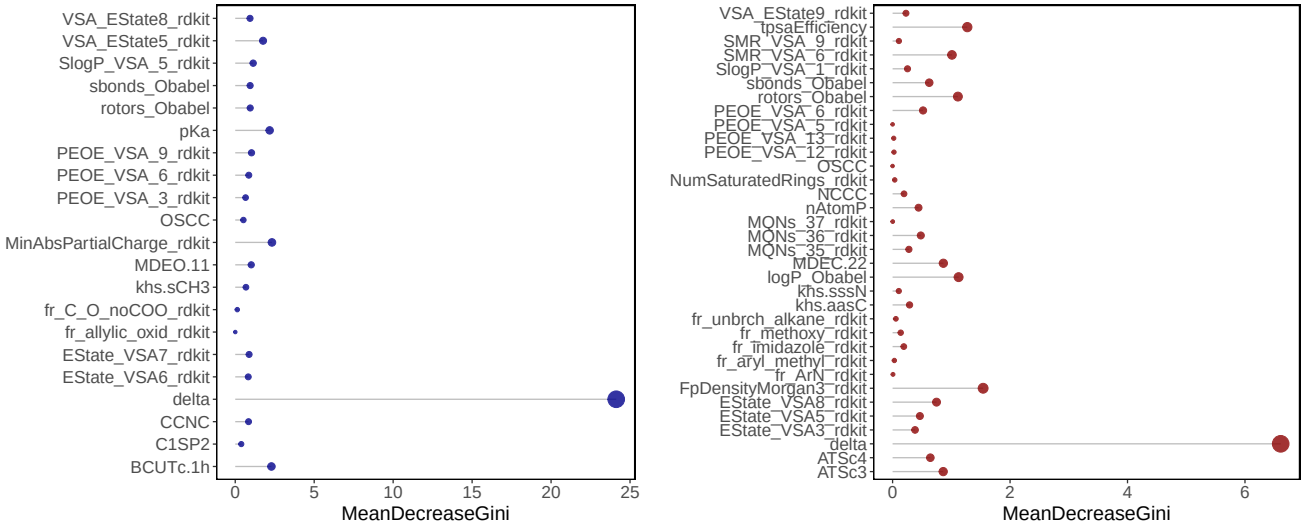


Figure 3. Importance of each descriptor from both acidic and basic models using the `meandecreaseGini` score from the `randomForest` library in *R*.

Logistic regression (LR), random forest (RF), and support vector machine with linear kernel (SVML) models were selected due to their consistently accurate performance in prior evaluations with both the test set and external datasets. Hyperparameter tuning was also explored in previous work.[31] Fig. 4 presents the classification results for the test set. All models produced relatively consistent results for the acidic dataset; however, the SVML model outperformed all other algorithms in the basic dataset. To mitigate inaccuracies in each individual model, their outputs were averaged, which resulted in more balanced outcomes for the acidic model. However, averaging the outputs of the basic models did not yield accurate results, primarily due to the low specificity of the basic LR model, which reduced the overall accuracy.

Additionally, it was observed that ROBERT faced challenges in developing an effective ML model for this classification problem. The ROBERT report was made publicly available in our GitHub repository, along with all relevant information to ensure reproducibility. The model was initially trained using AdaBoost, followed by training with different datasets and a 5-fold cross-validation. However, the model selected a 90:10 split for the training and test sets, which may have introduced a data imbalance due to the low proportion of positive output instances (requiring Eq. 2). Moreover, it is probable that the

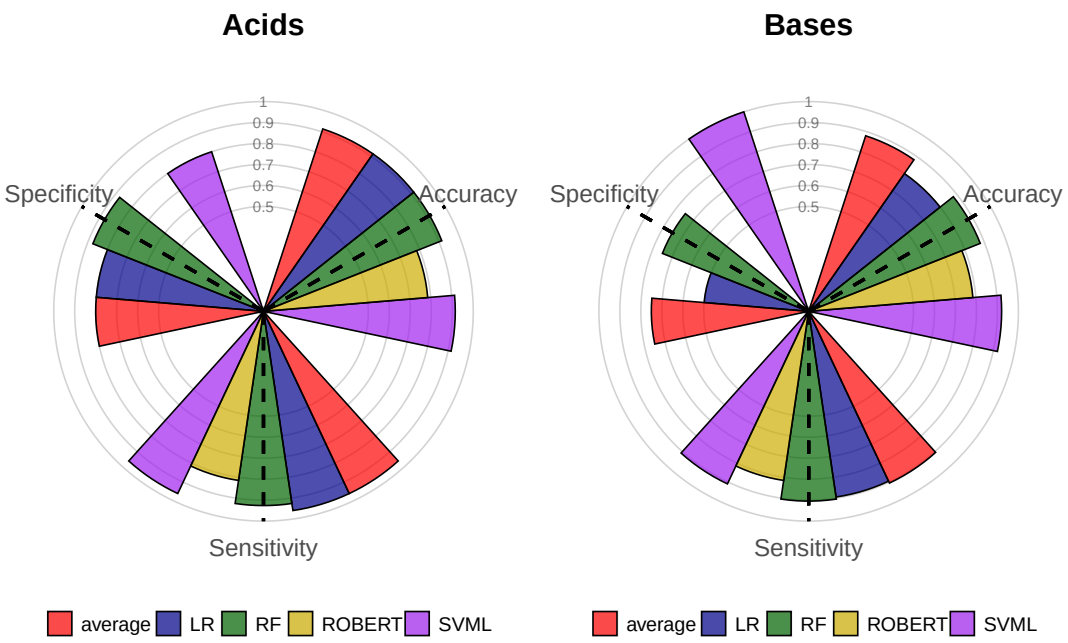


Figure 4. Accuracy, sensitivity, and specificity of each ML classification model after evaluating their performance with the test set. See the Supporting Information for the equation and Table S2 for the data.

descriptors generated by AQME lacked sufficient predictive power for this classification task. ROBERT includes a data curation tool that filters correlated, duplicated, and constant descriptors. Additionally, it employs Recursive Feature Elimination with Cross-Validation to showcase the most important descriptors.[41] During this process, it was observed that of the 215 original descriptors, only 3 exhibited significant capacity to differentiate between classifiers. Consequently, the ROBERT models displayed specificities of 0, indicating an inability to make accurate predictions for molecules requiring Eq. 2 (Fig. 4). This issue underscores the critical role of descriptor selection in our ML models. A combination of experimental features and those generated through chemoinformatics tools appears essential for accurate predictions, particularly considering the significance demonstrated by the `delta` descriptor (Fig. 3).

3.2 FAIR Google Colab Implementations of LiProS

LiProS is a straightforward tool, assisted by Google Colab, designed to facilitate the models developed for the scientific community. The initial step in this implementation involved selecting the most optimal models. For acidic molecules, the decision was made to implement the average of the three tested algorithms, as this approach addresses the sensitivity deficiencies of RF and the specificity issues of LR and SVML (Fig. 4, left). In the case of basic molecules, the SVML model was utilized due to its significantly superior performance compared to the others (Fig. 4, right).

To validate the functionality of LiProS, a test was conducted using the three previously studied alkaloids that had the highest Docking scores in the binding sites of the proteins *Trypanothione reductase* and *Cruzain*, which are essential proteins for the *Trypanosoma cruzi* parasites (Fig. 2). The pK_a values of the molecules were estimated using the predictor provided by ChemAxon, which has demonstrated reliable predictions for small molecules.[31] Additionally, LiProS was instructed to evaluate the models for these molecules at a pH of 7.4, due to the relevance of these molecules as potential human drugs.[43]

Table 1 presents the outputs generated by LiProS for each of the molecules. It is noteworthy that the molecules 2α,10bα-dihydroxy-9-O-dimethylhomolycorine and 10-Norneronine were labeled as basic, despite their ampholytic structure. Although their phenolic groups can deprotonate, alkaloids are notably basic molecules. LiProS estimated that these molecules require only Eq. 1 to accurately predict their $\log D_{7.4}$. This is possibly due to the fact that at this pH, the molecules exist predominantly in a neutral state (for both acidic and basic groups). Therefore, LiProS deduces (primarily through the `delta` descriptor) that the inclusion of P_I^{app} is not necessary to achieve an accurate prediction of their lipophilicity.

This example, applied in the context of structure-based drug design for Chagas disease, demonstrates the implementation of this tool under the FAIR principles.[35] It took less than 5 minutes to assess which lipophilicity formalism might be the most accurate for computationally predicting the $\log D_{7.4}$ of these molecules. In an active drug design process, the next step would

[1] pK_a values were estimated with ChemAxon's calculator.

Table 1. *Trypanothione reductase* and *Cruzain* inhibitor candidates analyzed with LiProS to asses the most appropriate $\log D_{\mathrm{pH}}$ formalism for each molecule.

ID	type	$\mathrm{p}K_{\mathrm{a}}$[1]	pH	LiProS prediction
Pancratinine A	base	6.6	7.4	"Use Eq. 1"
2α,10bα-dihydroxy-9-O-dimethylhomolycorine	base	6.6	7.4	"Use Eq. 1"
10-Norneronine	base	6.1	7.4	"Use Eq. 1"

involve finding accurate $\log P_{\mathrm{N}}$ and $\mathrm{p}K_{\mathrm{a}}$ predictors for this specific chemical space, followed by calculating $\log D_{7.4}$ using Eq. 1. These steps will provide critical information regarding the bioavailability of these compounds as potential drugs for this disease.

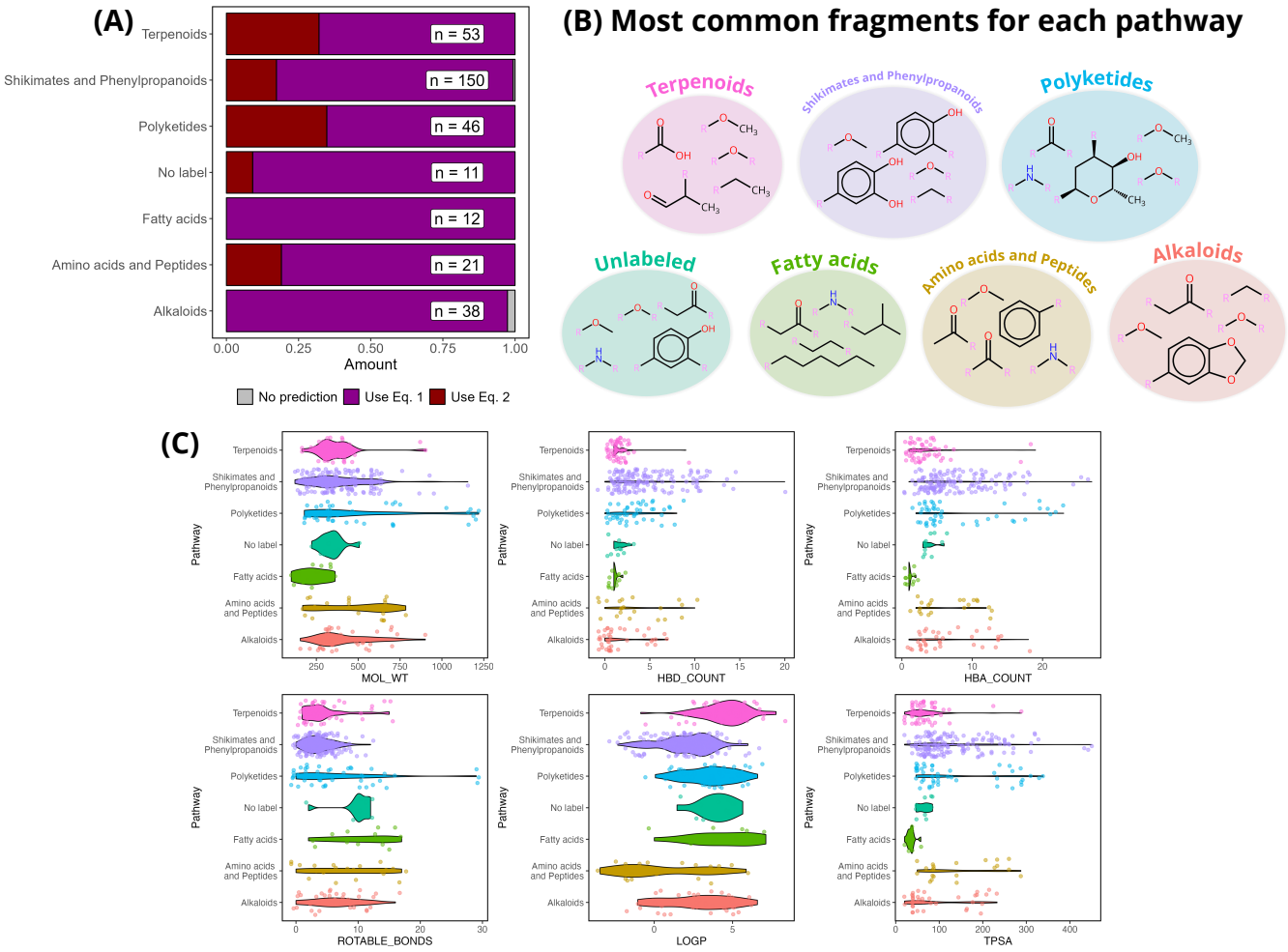


Figure 5. Study of the ionizable molecules of the NaProRE Database: (A) Barplot of LiProS's predictions for each natural product's pathway, (B) most common fragments or each NP pathways based on `RDKit`'s BRICS fragment decomposing algorithm, (C) visualization for the chemical space of the evaluated molecules for each pathway (**MOL-WT**: Molecular Weight, **HBD-COUNT**: Hydrogen-bond donors, **HBA-COUNT**: Hydrogen-bond acceptors, **ROTATABLE-BONDS**: Rotatable bonds, **LOGP**: calculated $\log P_{\mathrm{N}}$, **TPSA**: Topological polar surface area).

Similarly, the capability of LiPRoS to determine the most appropriate lipophilicity profile for a large number of molecules simultaneously was tested. For this purpose, we employed NaProRE-CR, a database of natural products reported in Costa Rica, which is part of the recently published Latin American database LaNaPDB.[44;45] Only the ionizable molecules of this

dataset were selected, utilizing the tools provided by Frolov *et al.* in the pIChemist package.[46] The pK_a values of the 323 remaining ionizable molecules were estimated using ChemAxon's predictor and analyzed using LiPRoS at a pH of 7.4. Fig. 5a presents the results obtained by LiPRoS for each pathway of the analyzed natural products. A significant proportion of Terpenoids, Shikimates and Phenylpropanoids, Polyketides, and Amino acids and Peptides were observed to fit better to Eq. 2. Consequently, the structural nature of each pathway was explored. The `BRICS.decompose` algorithm from `RDKit` was used to identify the five most frequent fragments in each pathway (additional fragments can be found in the `most-common-fragments.ipynb` notebook available in the GitHub repository). For instance, it was observed that Polyketides, Terpenoids, and Shikimates and Phenylpropanoids possess many fragments containing a high number of polar heteroatoms and ionizable groups. In these cases, the high presence of ionizable groups likely generates large δ values, which significantly influences the decision of the LiPRoS models. Additionally, an exploration of the chemical space of the analyzed molecules (Fig. 5c) revealed results consistent with previous studies, where pathways containing small molecules (molecular weight and TPSA), unsaturated compounds (low rotatable bonds), with many heteroatoms (HBA and HBD counts), and hydrophilic properties ($\log P_N \leq 0$) required Eq. 2 for calculating their $\log D_{7.4}$.[31]

In the group of Shikimates and Phenylpropanoids, and Alkaloids, two molecules were identified for which LiPRoS did not produce an output (**Germin D, 0** and **Borucoside, 262**, respectively). In these cases, a detailed analysis of the structures and descriptors was conducted. Firstly, the molecular masses of these compounds (634.455 $\frac{g}{mol}$ and 612.603 $\frac{g}{mol}$) were found to be significantly larger than the average molecular masses in the chemical space of the original dataset (< 400 $\frac{g}{mol}$). Also, similar trends were noted for HBA (> 5), HBD (> 2), and TPSA (> 100 Å).[31] Thus, these molecules are substantially larger than those in the dataset, making it likely that they fall outside the chemical space of the models. This phenomenon is further reflected when comparing the descriptor values for the acidic molecules in the LiPRoS dataset (Fig. S4), where the NaProRE-CR outliers showed significantly higher values in the majority of descriptors. The most notable case is with the `BCUTc.1h` descriptor. This descriptor is calculated by transforming the SMILES code into an adjacency matrix (ignoring hydrogens), where each value contains empirical atomic partial charges from the Gasteiger-Marsili method, with weights depending on the type of connectivity between atoms. The Eigenvalue for this matrix is then calculated, and the process is reiterated with a new matrix ordering to obtain new Eigenvalues.[47–49] The "1h" term in the descriptor indicates that only the highest Eigenvalue from all the calculations is sselected. In the case of these two outliers, values of $\sim -2.1 \times 10^9$ were obtained. This result suggests an error in the calculation of the Eigenvalues for these molecules, or even in the construction of the adjacency matrices. Fig. S5 illustrates the structures of these two outliers. It is observed that these molecules exhibit relatively complex connectivities, with numerous rings and complex stereochemistry. This complexity may have caused an error in the calculation of this descriptor using the `RCDK` package.

Through these two experiments (the Chagas molecules and the evaluation of NaProRE-CR), we assessed the ability of LiPRoS to determine the optimal lipophilicity profiles for both a small set of molecules and a significantly larger dataset. Under this framework, LiPRoS adheres to the following criteria:

- ***Findable:*** LiProS is readily discoverable and easy to implement. The Google Colab link is accessible directly from this manuscript. Additionally, the GitHub repository (*which also references the Colab link*) can be located through a Google search. Similarly, the data used to train the models are easy to find, as they are indexed both on GitHub and Zenodo.[36] Each dataset is accompanied by references to the source and the experimental technique used for measurement.

- ***Accessible:*** Google Colab is an open-access platform available to everyone, with the only requirement being a Google account. Furthermore, the GitHub repository is entirely open-access, and all documentation and scripts can be easily downloaded via a `!git clone` command in the terminal. These two features ensure that this protocol is open, free, and implementable. Moreover, authentication and authorization processes can be carried out, as the GitHub repository is fully public and allows users to leave comments and feedback, which can be incorporated into future versions.

- ***Interoperable:*** The LiProS Colab link enables multiple users to run the scripts (even simultaneously) without interfering with the original infrastructure. Additionally, this tool is designed so that the cells run almost automatically with minimal input. Consequently, LiProS can be used by individuals with limited programming knowledge. Furthermore, the Colab tool provides instructions throughout the code to guide users, ensuring the use of straightforward and accessible language in line with FAIR principles.

- ***Reusable:*** The tool and the format of the Google Colab sharing link allow users to modify the code as needed. However, upon refreshing the page, the original code is restored. LiProS generates results for small molecule sets in approximately 5 to 10 minutes. For larger datasets (e.g., > 100 molecules), the runtime increases with dataset size. Alternatively, users may opt for Colab's paid services, which provide access to GPUs with up to 500 compute units.[50] This time can also be reduced by running the script locally instead of through Google Colab. Our open acces databases and scripts facilitate free and open reproduction of the models, promoting continuous model improvement.

4 Conclusions

LiProS is a straightforward tool for predicting the most appropriate lipophilicity formalism to estimate the $logD_{\mathbf{pH}}$ of a molecule of interest. This tool facilitates the accurate prediction of lipophilicity for small molecules in the early stages of drug design, specifically for biocompatibility evaluation. Additionally, it can be employed by researchers in environmental chemistry, food chemistry, and medicinal chemistry, where understanding the lipophilicity of molecules is also crucial. LiProS was developed using improved ML classification models from previous works. Various models were trained on a well-curated and larger dataset, utilizing descriptors obtained from chemoinformatics packages in Python and R. The accuracy of the models was evaluated using a test set, which revealed that, for acidic compounds, an average of the three primary models (LR, RF, and SVML) provided more balanced and accurate results. In contrast, the basic SVML model exhibited significantly superior performance compared to the others. Consequently, these two models were implemented in LiProS, which is deployed as a simple Google Colab server. This server was meticulously developed following FAIR principles to ensure accessibility for the scientific community, even for those with limited programming expertise.

Finally, the LiProS Google Colab tool was applied to two cases: the evaluation three alkaloids with potential inhibitory activity against vital proteins of the causative agent of Chagas disease and a large-scale evaluation of the ionizable molecules of NaProRE-CR. Firstly, these examples demonstrated LiProS's capability to provide rapid results with minimal input, highlighting its applicability as a complement to the exhaustive work involved in drug design. Additionally, this application showcased LiProS's ability to discern the most suitable lipophilicity profile based on subtle structural differences, thereby underscoring the importance of the selected descriptors for the models.

In conclusion, LiProS has significant potential to facilitate and complement the prediction of lipophilicity for small, ionizable molecules. The scientific community working with this physicochemical property across various relevant fields of chemistry can greatly benefit from this tool.

5 Data availability

The curated dataset with experimental lipophilicity values can be found in Ref[36].

Further datasets, descriptors, scripts, and model training can be found in the repository https://github.com/cbio3lab/LiProS.git

The Google Colab server can be freely used with this link: https://colab.research.google.com/drive/1w9Vkqm4kIBQPn5AeSYnIwAs6vCTd8G7u?usp=sharing

6 Acknowledgements

We thank the Vice Chancellor for Research of the University of Costa Rica for their support via the research projects 115-C2-126 and 908-C3-610. We also thank Dr. Esteban Meneses and MSc. Fabricio Quirós Corella of the National Advanced Computing Colaboratory (CNCA) for their insights during this research. We also want to acknowledge the National High Technology Center (CeNAT) of Costa Rica for their support through their fellowship programmes (Beca CeNAT).

References

[1] lipophilicity. 2019; https://doi.org/10.1351/goldbook.LT06965.

[2] Waring, M. J. Lipophilicity in drug discovery. *Expert Opinion on Drug Discovery* **2010**, *5*, 235–248.

[3] Toulmin, A.; Wood, J. M.; Kenny, P. W. Toward Prediction of Alkane/Water Partition Coefficients. *Journal of Medicinal Chemistry* **2008**, *51*, 3720–3730, PMID: 18558667.

[4] Bannan, C. C.; Calabró, G.; Kyu, D. Y.; Mobley, D. L. Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water. *Journal of Chemical Theory and Computation* **2016**, *12*, 4015–4024, PMID: 27434695.

[5] Eksterowicz, J. E.; Miller, J. L.; Kollman, P. A. Calculation of Chloroform/Water Partition Coefficients for the N-Methylated Nucleic Acid Bases. *The Journal of Physical Chemistry B* **1997**, *101*, 10971–10975.

[6] Zamora, W. J.; Viayna, A.; Pinheiro, S.; Curutchet, C.; Bisbal, L.; Ruiz, R.; Ràfols, C.; Luque, F. J. Prediction of toluene/water partition coefficients in the SAMPL9 blind challenge: assessment of machine learning and IEF-PCM/MST continuum solvation models. *Physical Chemistry Chemical Physics* **2023**, *25*, 17952–17965.

[7] 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. *Pharmaceutical Research* **1998**, *15*, 209–215.

[8] Dumont, E.; Darracq, G.; Couvert, A.; Couriol, C.; Amrane, A.; Thomas, D.; Andrès, Y.; Le Cloirec, P. Determination of partition coefficients of three volatile organic compounds (dimethylsulphide, dimethyldisulphide and toluene) in water/silicone oil mixtures. *Chemical Engineering Journal* **2010**, *162*, 927–934, Citation Key: DUMONT2010927.

[9] 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 Medicinal Chemistry Letters* **2020**, *11*, 2421–2427.

[10] 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*, 1–9.

[11] Esser, H. O. A review of the correlation between physicochemical properties and bioaccumulation. *Pesticide Science* **1986**, *17*, 265–276.

[12] Lewis, D. F. V.; Jacobs, M. N.; Dickins, M. Compound lipophilicity for substrate binding to human P450s in drug metabolism. *Drug Discovery Today* **2004**, *9*, 530–537.

[13] 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 settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25.1. *Advanced Drug Delivery Reviews* **2001**, *46*, 3–26.

[14] Viayna, A.; Pinheiro, S.; Curutchet, C.; Luque, F. J.; Zamora, W. J. Prediction of n-octanol/water partition coefficients and acidity constants (pKa) in the SAMPL7 blind challenge with the IEFPCM-MST model. *Journal of Computer-Aided Molecular Design* **2021**, *35*, 803–811.

[15] Zamora, W. J.; Pinheiro, S.; German, K.; Ràfols, C.; Curutchet, C.; Luque, F. J. Prediction of the n-octanol/water partition coefficients in the SAMPL6 blind challenge from MST continuum solvation calculations. *Journal of Computer-Aided Molecular Design* **2020**, *34*, 443–451.

[16] Sun, Y.; Hou, T.; He, X.; Man, V. H.; Wang, J. Development and test of highly accurate endpoint free energy methods. 2: Prediction of logarithm of n-octanol–water partition coefficient (logP) for druglike molecules using MM-PBSA method. *Journal of Computational Chemistry* **2023**, *44*, 1300–1311.

[17] Ghose, A. K.; Crippen, G. M. Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity. *Journal of Computational Chemistry* **1986**, *7*, 565–577.

[18] Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 868–873.

[19] Lopez, K.; Pinheiro, S.; Zamora, W. J. Multiple linear regression models for predicting the n-octanol/water partition coefficients in the SAMPL7 blind challenge. *Journal of Computer-Aided Molecular Design* **2021**, *35*, 923–931.

[20] Kenney, D. H.; Paffenroth, R. C.; Timko, M. T.; Teixeira, A. R. Dimensionally reduced machine learning model for predicting single component octanol–water partition coefficients. *Journal of Cheminformatics* **2023**, *15*, 9.

[21] Avdeef, A. *Absorption and Drug Development.*; John Wiley Sons, 2012.

[22] Comer, J.; Tam, K. In *Pharmacokinetic Optimization in Drug Research*, 1st ed.; Testa, B., Van De Waterbeemd, H., Folkers, G., Guy, R., Eds.; Wiley, 2001; p 275–304.

[23] Liu, X.; Tu, M.; Kelly, R. S.; Chen, C.; Smith, B. J. Development of a computational approach to predict blood-brain barrier permeability. *Drug metabolism and disposition: the biological fate of chemicals* **2004**, *32*, 132–139.

[24] Colmenarejo, G. In silico prediction of drug-binding strengths to human serum albumin. *Medicinal Research Reviews* **2003**, *23*, 275–301.

[25] Iglesias, V.; Pintado-Grima, C.; Santos, J.; Fornt, M.; Ventura, S. In *Data Mining Techniques for the Life Sciences*; Carugo, O., Eisenhaber, F., Eds.; Methods in Molecular Biology; Springer US: New York, NY, 2022; p 197–211.

[26] Zamora, W. J.; De Souza, 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*, 236a.

[27] Porto, W. F.; Ferreira, K. C. V.; Ribeiro, S. M.; Franco, O. L. Sense the moment: A highly sensitive antimicrobial activity predictor based on hydrophobic moment. *Biochimica Et Biophysica Acta. General Subjects* **2022**, *1866*, 130070.

[28] Lupton, S. J.; Shappell, N. W.; Shelver, W. L.; Hakk, H. Distribution of Spiked Drugs between Milk Fat, Skim Milk, Whey, Curd, and Milk Protein Fractions: Expansion of Partitioning Models. *Journal of Agricultural and Food Chemistry* **2018**, *66*, 306–314.

[29] Niu, L.; Henneberger, L.; Huchthausen, J.; Krauss, M.; Ogefere, A.; Escher, B. I. pH-Dependent Partitioning of Ionizable Organic Chemicals between the Silicone Polymer Polydimethylsiloxane (PDMS) and Water. *ACS Environmental Au* **2022**, *2*, 253–262.

[30] Leeson, P. D.; Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature Reviews Drug Discovery* **2007**, *6*, 881–890.

[31] Bertsch, E.; Suñer, S.; Pinheiro, S.; Zamora, W. J. Critical Assessment of pH-Dependent Lipophilicity Profiles of Small Molecules: Which One Should We Use and In Which Cases? *ChemPhysChem* **2023**, *24*, e202300548.

[32] 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. *Journal of the American Chemical Society* **2001**, *123*, 10684–10690.

[33] 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.

[34] LogP and logD calculations | Chemaxon Docs. https://docs.chemaxon.com/display/docs/logp-and-logd-calculations.md.

[35] Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **2016**, *3*, 160018.

[36] Bertsch, E.; Suñer, S.; Pinheiro, S.; Zamora, W. J. Experimental n-octanol/water Partition/Distribution Coefficients Database for Small Molecules. 2024; https://zenodo.org/records/13208946.

[37] Willighagen, E. L. *Groovy Cheminformatics with the Chemistry Development Kit*, 2nd ed.; 2023.

[38] RDKit. https://www.rdkit.org/.

[39] O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **2011**, *3*, 33.

[40] 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*, 1–11.

[41] Dalmau, D.; Requena, J. V. A. ROBERT: Bridging the Gap between Machine Learning and Chemistry. *WIREs Computational Molecular Science* **2024**, *accpeted*.

[42] Alegre-Requena, J. V.; Sowndarya S. V., S.; Pérez-Soto, R.; Alturaifi, T. M.; Paton, R. S. AQME: Automated quantum mechanical environments for researchers and educators. *WIREs Computational Molecular Science* **2023**, *13*, e1663.

[43] Mora, D.; Schosinsky, F.; Castellón, J.; González, K.; Bastida, J.; Pinheiro, S.; Zamora, W. Finding for New Drugs in Nature for Neglected Tropical Diseases: In Silico Study of Homolycorine Alkaloids Against Chagas Disease. 2023 IEEE 5th International Conference on BioInspired Processing (BIP). San Carlos, Alajuela, Costa Rica, 2023; p 1–7.

[44] Gómez-García, A.; Jiménez, D. A. A.; Zamora, W. J.; Barazorda-Ccahuana, H. L.; Chávez-Fumagalli, M. ; Valli, M.; Andricopulo, A. D.; Bolzani, V. d. S.; Olmedo, D. A.; Solís, P. N.; Núñez, M. J.; Rodríguez Pérez, J. R.; Valencia Sánchez, H. A.; Cortés Hernández, H. F.; Medina-Franco, J. L. Navigating the Chemical Space and Chemical Multiverse of a Unified Latin American Natural Product Database: LANaPDB. *Pharmaceuticals* **2023**, *16*, 1388.

[45] Zamora, W. J.; Pinheiro, S.; Acuña, D.; Zuñiga, J. NAPRORE-CR (NAtural PROducts REpository - Costa Rica). 2023; https://zenodo.org/

records/7858102.

[46] Frolov, A. I.; Chankeshwara, S. V.; Abdulkarim, Z.; Ghiandoni, G. M. pIChemiStFree Tool for the Calculation of Isoelectric Points of Modified Peptides. *Journal of Chemical Information and Modeling* **2023**, *63*, 187–196.

[47] Burden, F. R. Molecular identification number for substructure searches. *Journal of Chemical Information and Computer Sciences* **1989**, *29*, 225–227.

[48] Pearlman, R. S.; Smith, K. Novel software tools for chemical diversity. *Perspectives in Drug Discovery and Design* **1998**, *9*, 339–353.

[49] Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *Journal of Chemical Information and Computer Sciences* **1999**, *39*, 28–35.

[50] Colab paid services pricing. https://colab.research.google.com/signup?authuser=0&utm_source=notebook_settings&utm_medium=link&utm_campaign=premium_gpu_selector.

7 Supporting Information

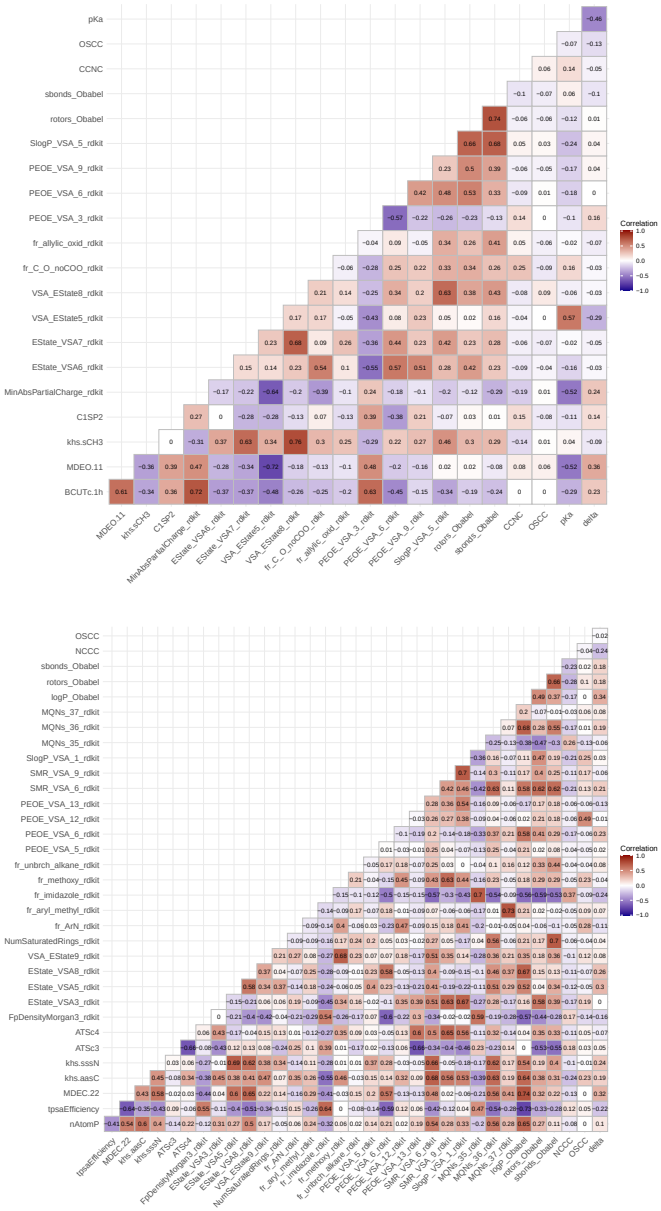


Figure S1. Correlations between de selected descriptors for the (*top*) acidic and (*bottom*) basic classification models.

Equations used to determine the accuracy, sensitivity, and specificity:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{3}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{4}$$

$$\text{Accuracy} = \frac{TP+TN}{n} \tag{5}$$

Table S1. Selected descriptors for the acidic and basic ML classification models after an univariate feature selection by performing a two-tailed Welch's *t*-test.

Descriptor	p-value (Welch's t-test)	Source
Basic dataset		
nAtomP	0.010262	RCDK
tpsaEfficiency	0.010373	RCDK
MDEC.22	0.004225	RCDK
khs.aasC	0.001545	RCDK
khs.sssN	0.019961	RCDK
ATSc3	0.000483	RCDK
ATSc4	0.046884	RCDK
FpDensityMorgan3_rdkit	0.004136	RDKit
EState_VSA3_rdkit	0.031416	RDKit
EState_VSA5_rdkit	0.030389	RDKit
EState_VSA8_rdkit	0.001241	RDKit
VSA_EState9_rdkit	0.001128	RDKit
NumSaturatedRings_rdkit	0.027731	RDKit
fr_ArN_rdkit	0.001467	RDKit
fr_aryl_methyl_rdkit	0.001317	RDKit
fr_imidazole_rdkit	0.013034	RDKit
fr_methoxy_rdkit	0.039099	RDKit
fr_unbrch_alkane_rdkit	0.017878	RDKit
PEOE_VSA_5_rdkit	0.007139	RDKit
PEOE_VSA_6_rdkit	0.037977	RDKit
PEOE_VSA_12_rdkit	0.000783	RDKit
PEOE_VSA_13_rdkit	0.000715	RDKit
SMR_VSA_6_rdkit	7.31E-05	RDKit
SMR_VSA_9_rdkit	0.018222	RDKit
SlogP_VSA_1_rdkit	0.046538	RDKit
MQNs_35_rdkit	0.007225	RDKit
MQNs_36_rdkit	0.0489	RDKit
MQNs_37_rdkit	0.044968	RDKit
logP_Obabel	0.008719	OpenBabel
rotors_Obabel	0.018195	OpenBabel
sbonds_Obabel	0.021024	OpenBabel
NCCC	0.044968	ConjugaR
OSCC	0.033613	ConjugaR
delta	1.14E-07	Experimental
Acidic dataset		
BCUTc.1h	0.005802	RCDK
MDEO.11	0.008455	RCDK
khs.sCH3	0.001195	RCDK
C1SP2	0.032561	RCDK
MinAbsPartialCharge_rdkit	0.005477	RDKit
EState_VSA6_rdkit	0.00219	RDKit
EState_VSA7_rdkit	0.002641	RDKit
VSA_EState5_rdkit	0.028259	RDKit
VSA_EState8_rdkit	2.40E-05	RDKit
fr_C_O_noCOO_rdkit	0.029608	RDKit
fr_allylic_oxid_rdkit	0.00239	RDKit
PEOE_VSA_3_rdkit	0.021602	RDKit
PEOE_VSA_6_rdkit	0.001216	RDKit
PEOE_VSA_9_rdkit	0.04263	RDKit
SlogP_VSA_5_rdkit	6.55E-05	RDKit
rotors_Obabel	0.003614	OpenBabel
sbonds_Obabel	0.021345	OpenBabel
CCNC	0.03202	ConjugaR
OSCC	0.027511	ConjugaR
pKa	0.014955	Experimental
delta	8.91E-18	Experimental

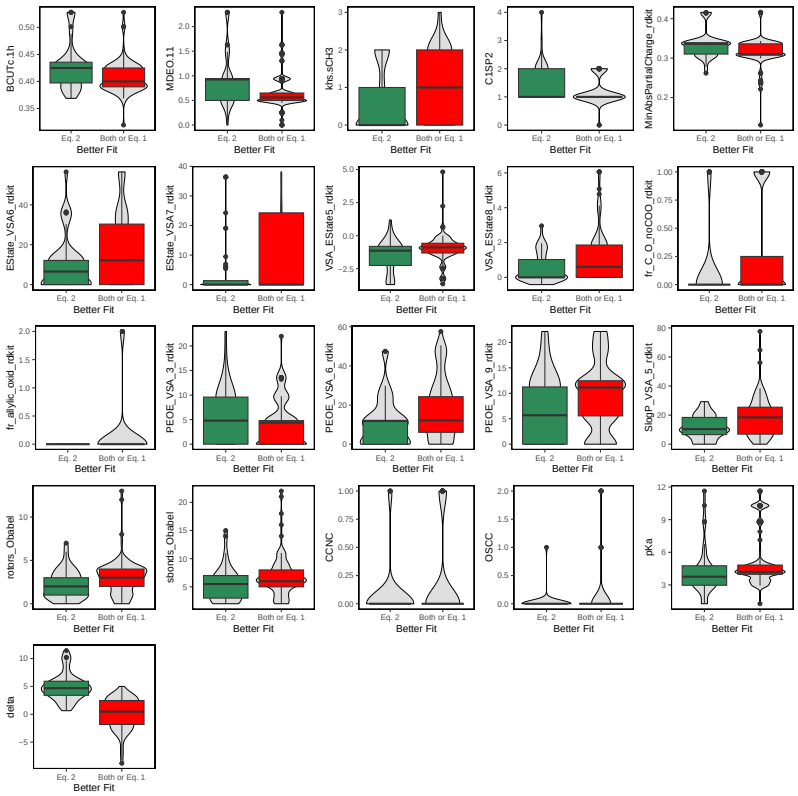


Figure S2. Violin plot of the distribution of values for each classification for the selected descriptor from the acidic dataset.

Table S2. Classification results for each ML model.

Result	LR	RF	SVML	Average
Acidic dataset				
True Positive	8	7	8	8
False Negative	2	1	2	2
False Positive	1	2	1	1
True Negative	24	25	24	24
Basic dataset				
True Positive	3	3	3	3
False Negative	3	1	0	1
False Positive	2	2	2	2
True Negative	17	19	20	19

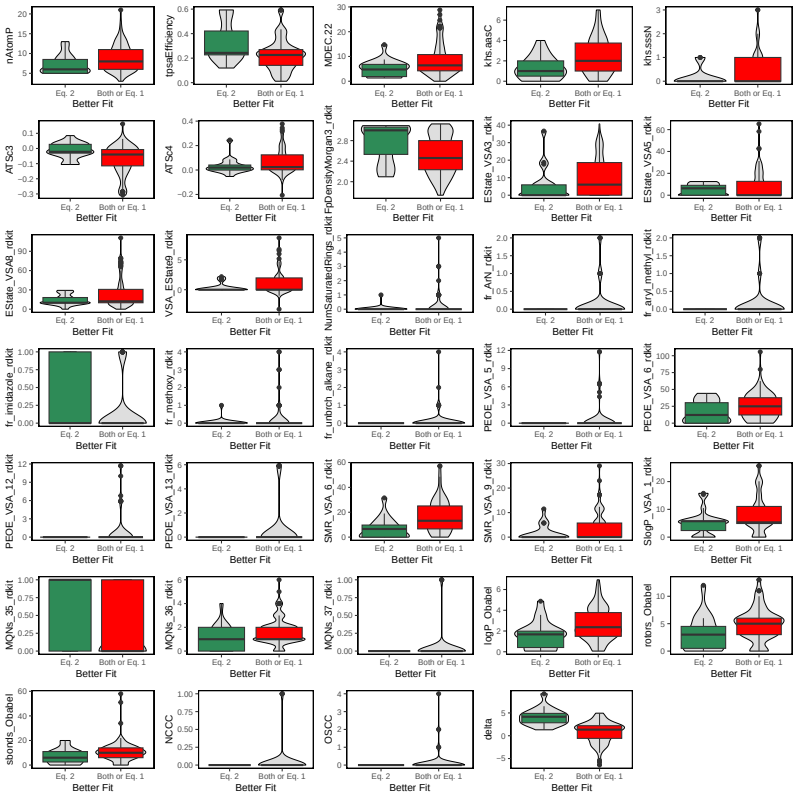


Figure S3. Violin plot of the distribution of values for each classification for the selected descriptor from the basic dataset.

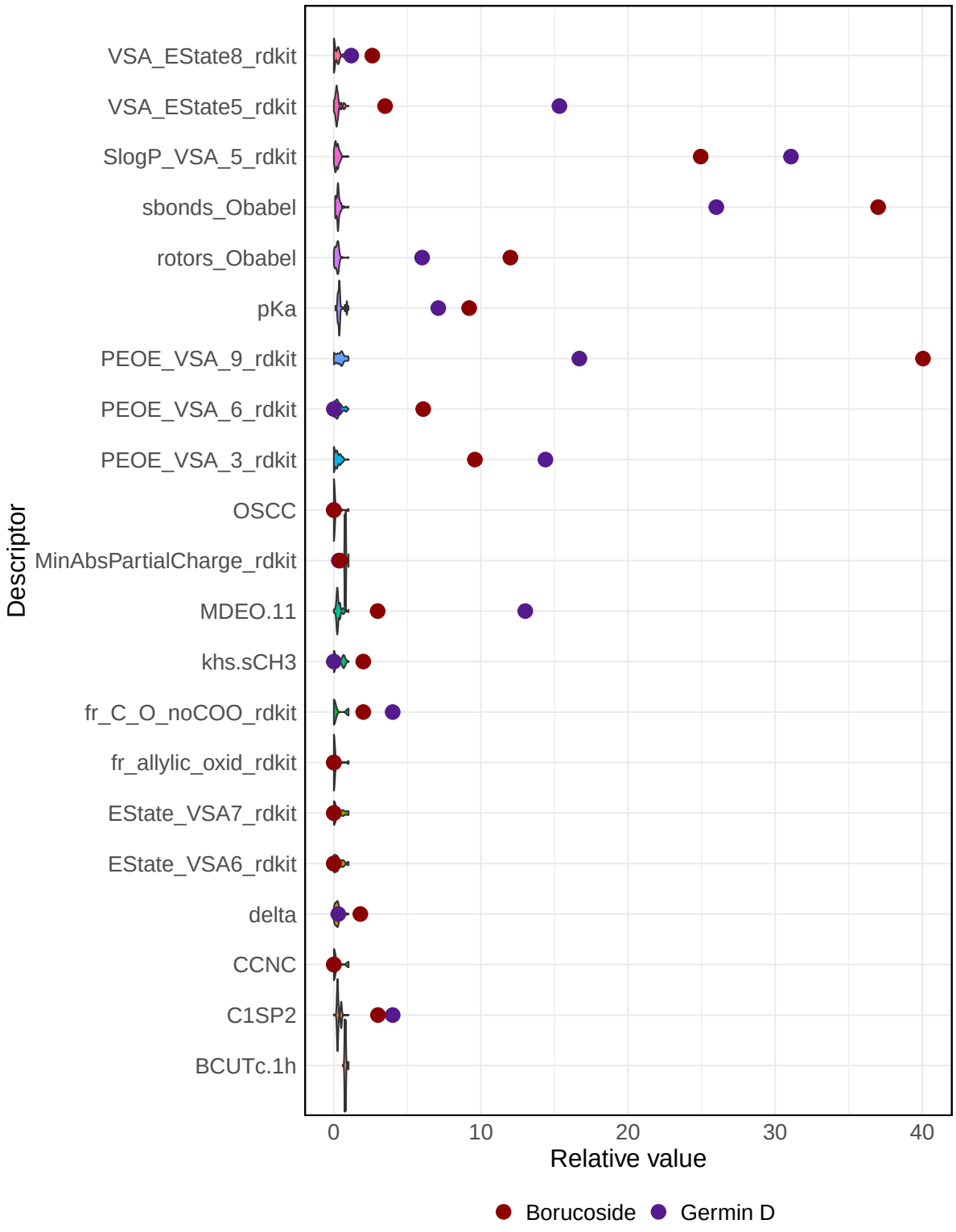


Figure S4. Visualization of the acidic descriptors for the training dataset in comparison to the NaProRE-CR outliers that did not display an output with LiProS. Relative values were calculated by dividing the absolute value of each data entry with the maximum value for each descriptor. BCUTc.1h values for the outliers are not displayed due to the large divergence of these values compared to the rest of the dataset.

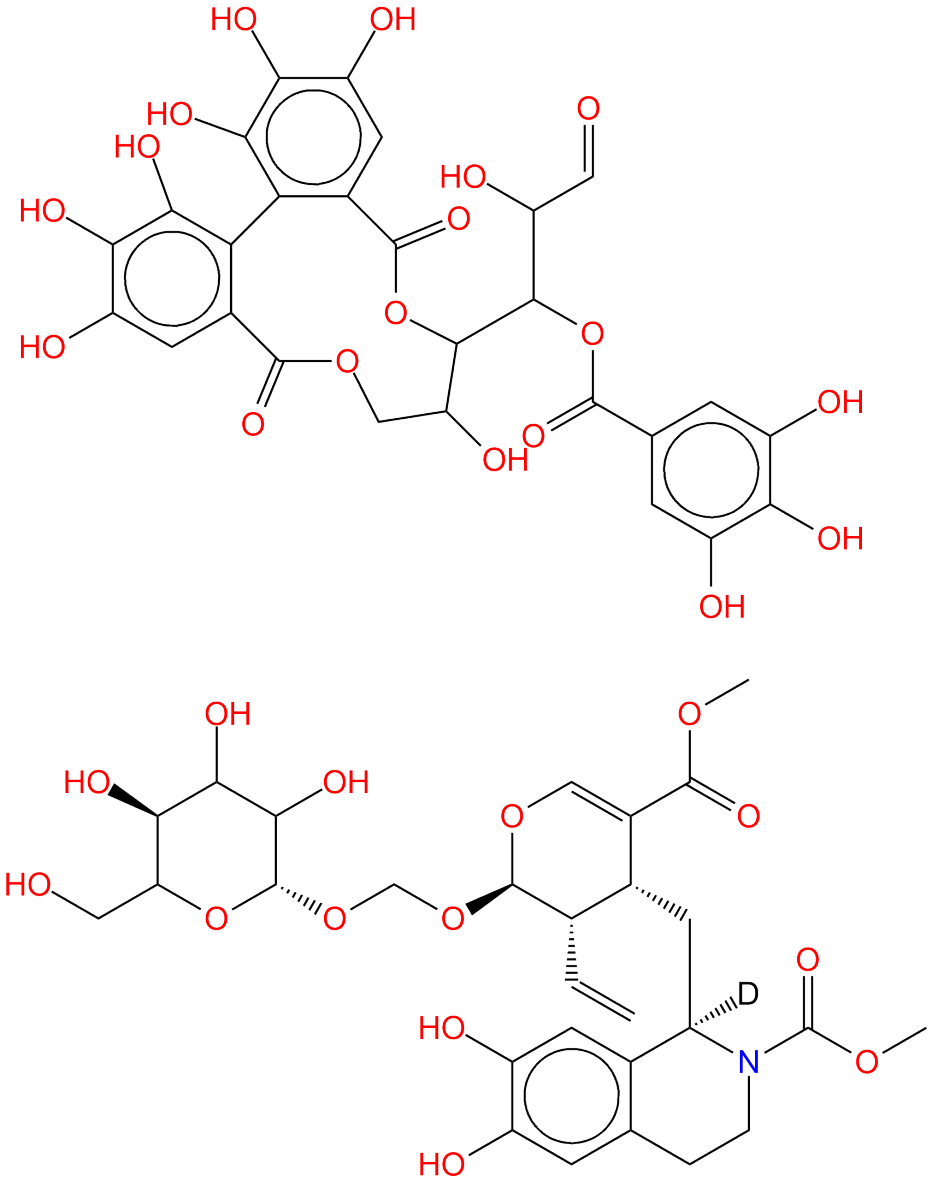


Figure S5. Molecular structure of *(up)* Borucoside and *(down)* Germin D.