Prediction of Toluene/Water Partition Coefficient in the SAMPL9 Blind Challenge: Assessment of Machine Learning and IEF-PCM/MST Continuum Solvation Models

William J. Zamora[1,2,*], Antonio Viayna[3,4,5], Silvana Pinheiro[1], Carles Curutchet[5,6], Laia Bisbal[4,7], Rebeca Ruiz[8], Clara Ràfols[4,7,*], F. Javier Luque[3,4,5,*]

[1] CBio[3] Laboratory, School of Chemistry, Laboratory of Computational Toxicology and Biological Testing Laboratory (LEBi), University of Costa Rica, San Pedro, San José, Costa Rica.

[2] Advanced Computing Lab (CNCA), National High Technology Center (CeNAT), Pavas, San José, Costa Rica

[3] Departament de Nutrició, Ciències de l'Alimentació i Gastronomia, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona (UB), Av. Prat de la Riba 171, 08921 Santa Coloma de Gramenet, Spain.

[4] Institut de Biomedicina (IBUB), Universitat de Barcelona (UB), Barcelona, Spain.

[5] Institut de Química Teòrica i Computacional (IQTC-UB), Universitat de Barcelona (UB), Barcelona, Spain.

[6] Departament de Farmàcia i Tecnologia Farmacèutica, i Fisicoquímica, Facultat de Farmàcia i Ciències de l'Alimentació, Universitat de Barcelona (UB), Av. Joan XXIII 27-31, 08028, Barcelona, Spain.

[7] Departament d'Enginyeria Química i Química Analítica, Universitat de Barcelona (UB), Martí i Franquès 1-11, 08028 Barcelona, Spain

[8] Pion Inc., Forest Row Business Park, Forest Row RH18 5DW, UK.

* Corresponding author: william.zamoraramirez@ucr.ac.cr (WJZ), crafols@ub.edu (CR), fjluque@ub.edu (FJL)

ORCID

WJZ: 0000-0003-4029-4528
AV: 0000-0002-2112-5828
SP: 0000-0002-6909-1129
CC: 0000-0002-0070-1208
RR: 0000-0002-0648-3176
CR: 0000-0001-7811-986X
FJL: 0000-0002-8049-3567

Abstract

In recent years the use of partition systems other than the widely used biphasic *n*-octanol/water has received increased attention to gain insight into the molecular features that dictate the lipophilicity of compounds. Thus, the difference between *n*-octanol/water and toluene/water partition coefficients has proven to be a valuable descriptor to study the propensity of molecules to form intramolecular hydrogen bonds and exhibit chameleon-like properties that modulate solubility and permeability. In this context, this study reports the experimental toluene/water partition coefficients ($\log P_{tol/w}$) for a series of 16 drugs that were selected as an external test set in the framework of the Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) blind challenge. This external set has been used by the computational community to calibrate their methods in the current edition (SAMPL9) of this contest. Furthermore, the study also investigates the performance of two computational strategies for the prediction of $\log P_{tol/w}$. The first relies on the development of two machine learning (ML) models, which are built up by combining the selection of 11 molecular descriptors in conjunction with either multiple linear regression (MLR) and random forest regression (RFR) models to target a dataset of 252 experimental $\log P_{tol/w}$ values. The second consists of the parametrization of the IEF-PCM/MST continuum solvation model from B3LYP/6-31G(d) calculations to predict the solvation free energies of 163 compounds in toluene and benzene. The performance of the ML and IEF-PCM/MST models has been calibrated against external test sets, including the compounds that define the SAMPL9 $\log P_{tol/w}$ challenge. The results are used to discuss the merits and weaknesses of the two computational approaches.

Introduction

Lipophilicity is a key physicochemical parameter in early-stage drug discovery due to its impact on the biodistribution of compounds and to the influence on diverse properties such as solubility, bioavailability and metabolic stability [1–5]. Lipophilicity is generally measured as the partition coefficient of a neutral compound between *n*-octanol and water ($\log P_{o/w}$). Numerous computational [6–10] and experimental [11–13] approaches have been developed for the prediction of $\log P_{o/w}$, whose relevance can be illustrated by the widely used Lipinski's rule-of-five [14], where the partition coefficient is a crucial component for determining the drug-likeness of compounds for oral delivery. Furthermore, lipophilicity continues to be a key property for the evaluation of new drug candidates such as PROTACs, which are considered to have physicochemical profiles beyond Lipinski's rule [15–17].

Although $\log P_{o/w}$ is an essential property for the prediction of the ADME-Tox properties of new drugs, the *n*-octanol/water system is a simple surrogate to explore the partition of compounds in biological environments. Besides *n*-octanol, which is characterized by combining the apolar aliphatic chain with the hydrogen-bond donor/acceptor function of the hydroxyl group, other organic solvents have been proposed as alternatives to imitate different physiological cell barriers, including chloroform/water, alkane/water, 1,2-dichloroethane/water, dibutyl ether/water, toluene/water and propylene glycol dipelargonate/water [18–24]. In this context, the difference between two partition coefficients using common organic solvents (e.g., *n*-octanol and alkanes) and water as polar phase has been employed to study the solubility of drugs [25]. Furthermore, the experimental alkane/water and toluene/water partition systems have given insights into the 'chameleonic' properties of drugs whose chemical space is beyond of the metrics of the rule-of-five, although experimental limitations that fall on the low solubility of compounds

87 in alkane-type solvents may limit their application [26–28]. As an example, it has been shown
88 that macrocyclic drugs populate significantly less polar and more compact conformational
89 ensembles in an apolar than in a polar environment, which enables to balance aqueous solubility
90 and cell permeability [22]. In turn, understanding the role of intramolecular hydrogen bonds and
91 their effect on lipophilicity can be valuable to optimize solubility and permeability of small
92 molecules, cyclic peptides and macrocycles [29].
93 Despite the increasing relevance of the toluene/water partition coefficient ($\log P_{tol/w}$) to modulate
94 the physicochemical properties of compounds beyond the rule-of-five chemical space, a limited
95 number of computational tools has been developed for the prediction of $\log P_{tol/w}$. To the best of
96 our knowledge, the only free empirical method available to estimate the $\log P_{tol/w}$ is Abraham's
97 partition equation based on the linear solvation energy relationship (LSER) method, which relies
98 on the use of five solute descriptors that can be obtained either experimentally [25,30] or
99 theoretically from databases [31]. On the other hand, quantum mechanical self-consistent
100 reaction field (QM-SCRF) models, where the solute is treated at the QM level and the solvent is
101 represented as a continuum polarizable medium characterized by suitable macroscopic
102 properties, have been shown to be a powerful approach for the calculation of the solvation free
103 energy in a variety of solvents [32-35], often combined with discrete treatment of solvent
104 molecules [36]. Besides computational efficiency, QM-SCRF methods are valuable to gain
105 insight into the effect of solvation on the geometrical and electronic properties of solutes. To the
106 best of our knowledge, specific parametrizations for the description of solvation effects in
107 toluene were implemented in the Solvation Model based on Density (SMD; [38]) and in the
108 Conductor-like Screening Model for Realistic Solvation (COSMO-RS; [39]) models.

Given the potential impact of the log$P_{tol/w}$ in drug discovery, this study reports the experimental determination for a set of 16 compounds chosen as external validation set within the framework of the Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) blind challenge. Previous SAMPL editions have focused their attention on the prediction of the log$P_{o/w}$ (and log$D$) for molecular systems selected in order to calibrate the merits and capabilities of contemporary methods [10, 39-44]. Following this spirit, this study describes the experimental determination of the log$P_{tol/w}$ values determined for the external test set, which will be used for the statistical assessment of computational approaches for the prediction of log$P_{tol/w}$ in the current SAMPL edition (SAMPL9) [45]. Furthermore, the manuscript reports the parametrization of two prediction models. This first consists of a machine learning (ML) model developed by exploiting 252 unique experimental log$P_{tol/w}$ values, and the second involves the parametrization of the B3LYP/6-31G(d) version of the integral IEF-PCM/MST continuum solvation model [46] in toluene and benzene.

Materials and methods

Experimental determination of the toluene/water partition coefficients

16 compounds with acid-base properties were selected for determining their experimental log$P_{tol/w}$ values (Fig. 1). These compounds were obtained from Sigma-Aldrich (purity ≥ 98%). The partition solvent (toluene) was also obtained from Sigma-Aldrich (ACS reagent, purity ≥ 99.5%).

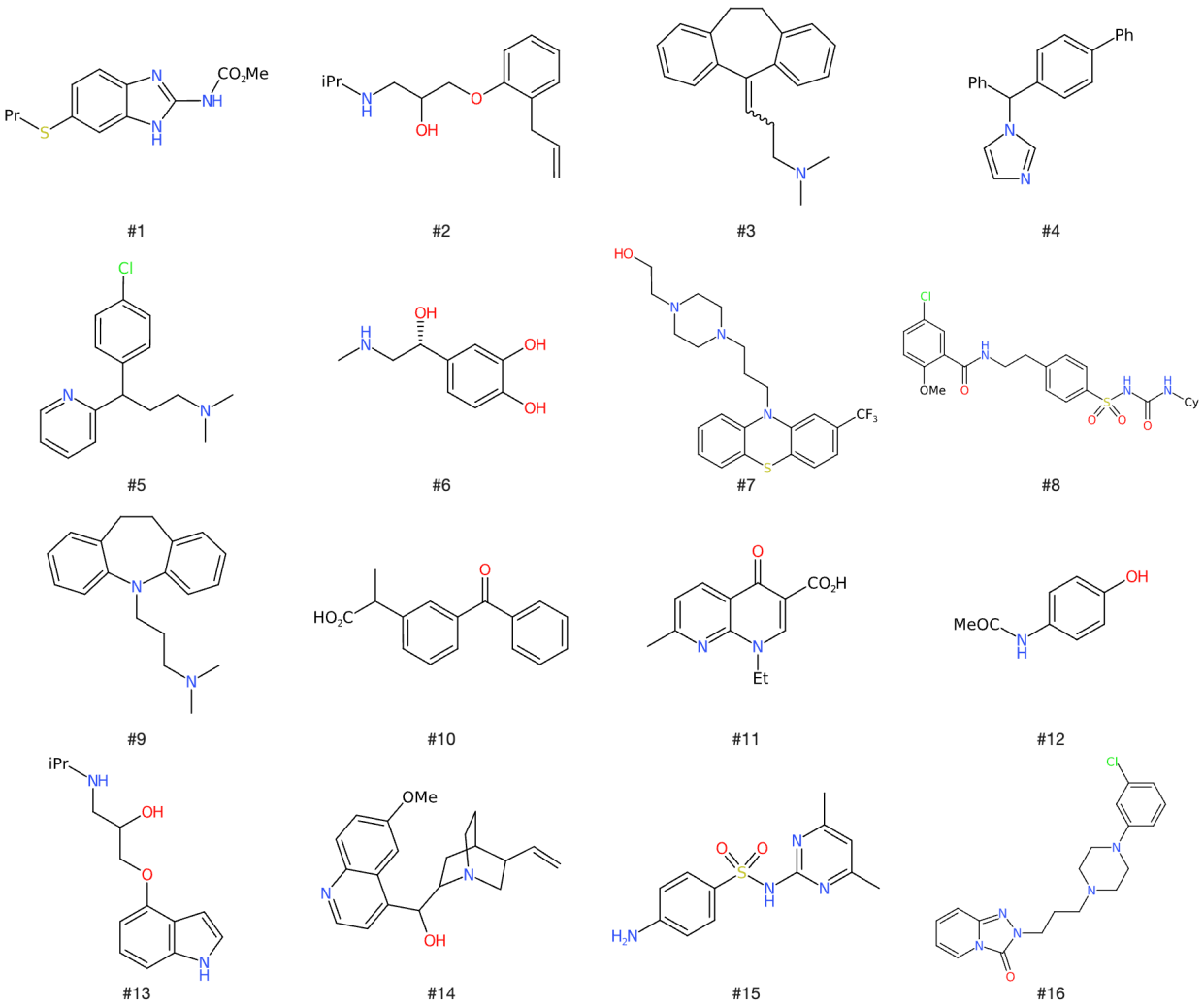


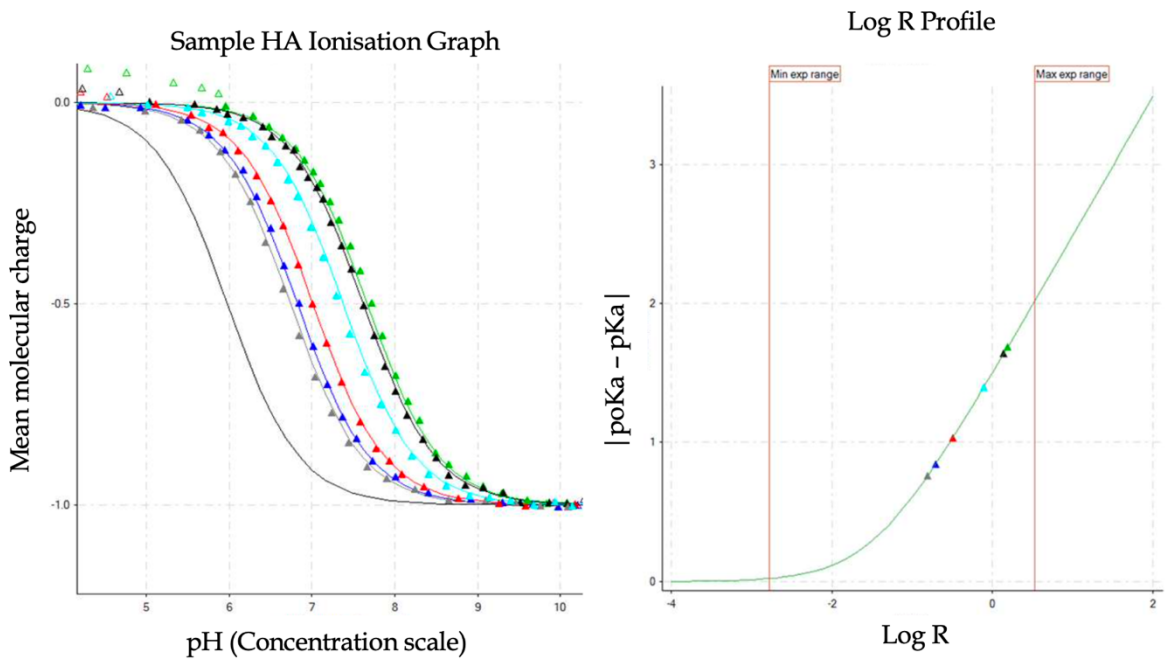
Fig. 1 Chemical structures of the drugs used as an external set in this work.

The log$P_{tol/w}$ values were obtained by sample titrations using Pion SiriusT3 (Pion Inc.) and Sirius D-PAS & GLpKa (Sirius Analytical Instruments Ltd.) in the presence of various volumes of the partitioning solvent. The log$P_{solv/w}$ was determined measuring the shift between the pK_a previously determined in aqueous media (KCl = 0.15 M) and the apparent p$_o K_a$ (pK_a measured in presence of toluene) at several sequential partition ratios (see Supporting Information Table S1) using the mass and balance equations. All measurements were made at 298 K, under an inert gas

(argon or nitrogen) atmosphere, and at least three titrations were carried out for each compound [47].

In cases where the compound had a basic natural state, the solution was pre-acidified to an appropriate pH and titrated with KOH towards basic pH values. On the other hand, if the compound had an acid natural state, the titration was started at high pH values, and it progressed towards low pHs. This process was repeated several times with different toluene/water partition volumes for every compound.

The initial partition volumes were selected according to the default volumes defined by the instruments (see Supporting Information Table S1), and the following partitions were optimized using the log*R* profile (log10 of toluene to water volumetric ratio) provided by SiriusT3Refine software. From the selected titrations, the log$P_{tol/w}$ can be refined and evaluated as an average using the SiriusT3Refine software [48,49]. An example of pH-metric log$P_{tol/w}$ determination results for nalidixic acid is shown in Fig. 2, where the Bjerrum plot shows the average number of bound protons in relation to the pH.



157

158 **Fig. 2** Experimental (left) Bjerrum plot curves and (right) log*R* profiles obtained in the procedure
159 for a moderately hydrophobic substance (nalidixic acid). Vertical lines in the right plot denote
160 minimum and maximum experimental ranges.

161

162 **Machine Learning models to determine the toluene/water partition coefficient**

163 An initial set of 307 log$P_{tol/w}$ values taken from different studies in the literature [24,28,50] was
164 compiled to calibrate the performance of the ML approach. With the only exception of 4-
165 hydroxybenzoic acid, where the difference between the reported values is close to 0.8 log*P* units,
166 slightly different log$P_{tol/w}$ values (i.e., the difference is generally < 0.2 logP units) were found for
167 certain compounds in the data collected from these studies. In these cases, the average value was
168 annotated, thus obtaining a final dataset of 252 unique molecules whose SMILES were
169 canonicalized using Open Babel [51] (see Supporting Information DatabaseTol_feb23.csv). Fig.
170 3 shows the distribution of data according to the classification of the compounds in their
171 acid/base character. Almost 50% of the compounds are neutral and there is a similar balance

between acidic and basic compounds. Most of the compounds have a molecular weight lower than 200, although values close to 500 are reached in few instances. On the other hand, the molecules exhibit low polarity, as the number of HB acceptors is generally ≤ 2 and the number of HB donors is ≤ 1. Finally, they exhibit limited conformational flexibility, as most of the compounds have a number of rotatable bonds ≤ 2.

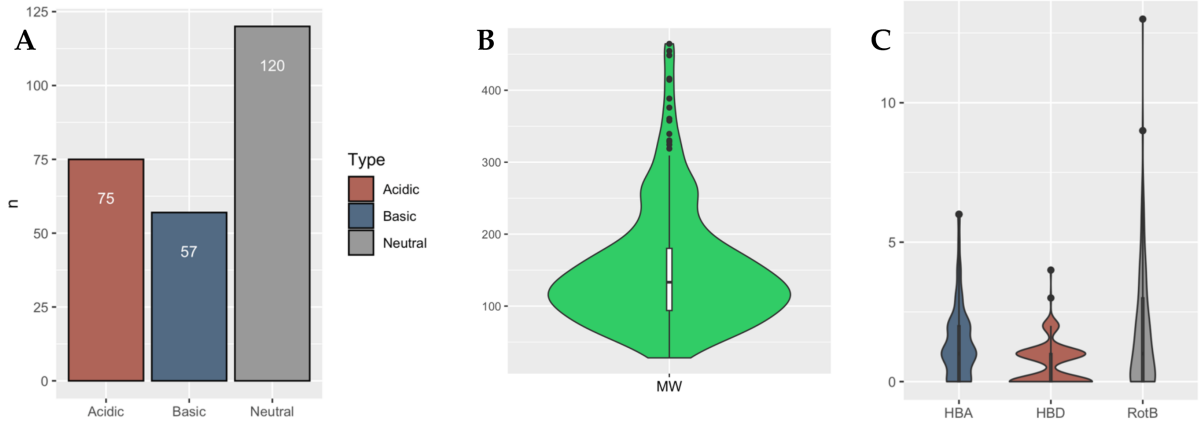


Fig. 3 Distribution of the database of 252 small molecules based on (**A**) the classification in neutral, acidic, and basic compounds, (**B**) molecular weight (MW, Da), and (**C**) number of hydrogen bond acceptors (HBA) and donors (HBD), and number of rotatable bonds (RotB).

A total of 453 molecular descriptors were calculated using the open-source Java framework for cheminformatics Chemistry Development Kit (CDK) [52] and the R environment from the canonical SMILES of the compounds. The descriptors comprise five categories [53]: i) descriptors that reflect the molecular composition of a compound without any information about its molecular geometry (e.g., number of atoms, bonds, rings, etc.), ii) topological properties obtained from a molecular graph (usually H-depleted, e.g., conformationally independent 2D-descriptors), iii) geometric parameters derived from the 3D structure of the molecule (e.g., radius of gyration, molecular eccentricity, 3D polar surface area), iv*)* electronic properties associated

with the electron density distribution in the molecule (e.g., atomic charge, dipole moment, energy of frontier orbitals), and v) hybrid descriptors falling into two or more of the categories mentioned before.

In addition, 15 descriptors based on a *n* conjugated system with $n = 4$ ($n = \pi$ electrons) were also included to account for a molecular count descriptor that could be relevant for the partition in an aromatic solvent as toluene (Fig. 4; see Supporting Information cjsystems.csv and conjugatedset.R). They were derived based on the ChemmineR package in R and the sdf files generated from the SMILES code of the compounds.

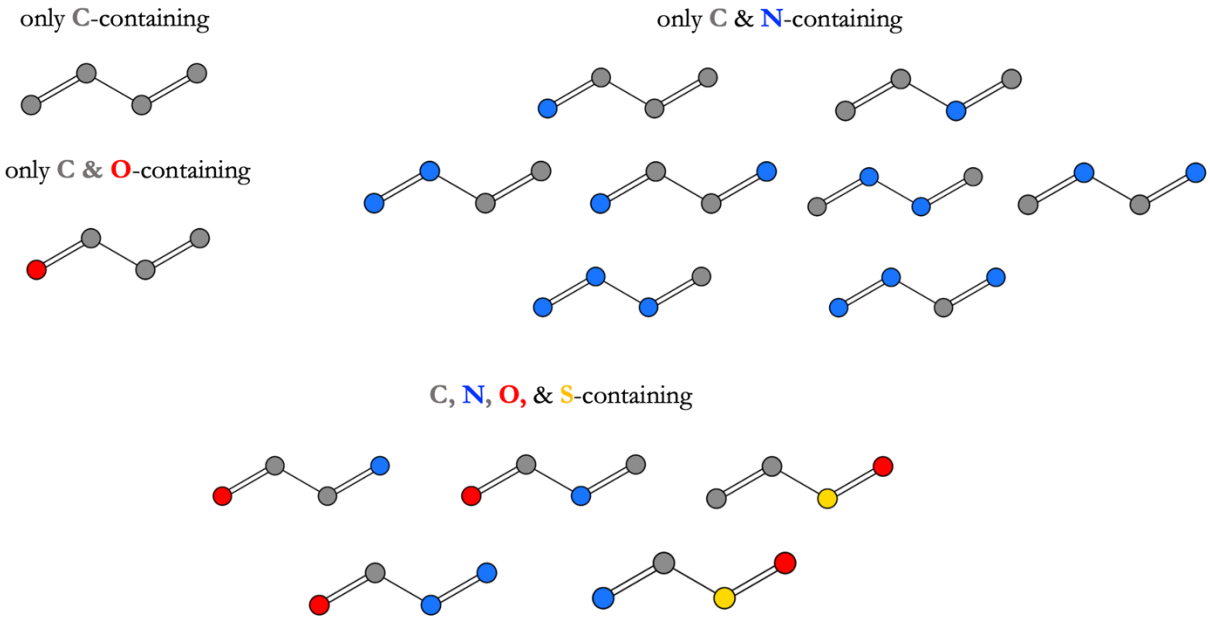


Fig. 4 Representation of *n* conjugated systems with n = 4 ($n = \pi$ electrons) detected in open-chain and cyclic compounds. The atoms forming the conjugated system are depicted in their respective colors, carbon (grey), nitrogen (blue), oxygen (red), and sulfur (yellow). The models were designed using Smarts Plus [35].

Multiple linear regression (MLR) and random forest regression (RFR) were chosen to explore the predictive log$P_{tol/w}$ models, whose performance and statistical analysis were determined using the R software environment. In order to internally validate the models, 85% of the data (214 compounds; see Supporting Information DatabaseTol_feb23.csv) were used for the training set, and 15% (38 compounds; see Supporting Information DatabaseTol_feb23.csv) for the test set.

The whole set of 468 descriptors (see above) was first filtered out by eliminating highly correlated descriptors (i.e., with a determination coefficient $r^2 > 0.70$). The remaining 79 descriptors were then used to perform an exhaustive selection analysis with the R package called *Leaps* [55] to choose the most significant subset of dependent variables for the MLR model. Accordingly, a final subset of 11 predictors (see Table 1) was selected, further checked for mutual correlations (see Supporting Information Fig. S1), and finally used for the MLR and RFR models.

Table 1 List of the most influential predictors used in this study for the regression models and their range of values for the set of compounds included in the training set.

Descriptor	Type	Definition	Range in training set (min, max)
Halogens	Topological CDK descriptor	Count of halogen atoms	(0, 4)
CCCO	Descriptor of conjugated systems	Count of conjugated C=C-C=O (Fig. 4)	(0, 4)
BCUTc.1l	Hybrid CDK descriptor	Eigenvalue-based descriptor of the lowest partial charge weighted BCUTS (details in Supporting Information Fig. S3) [57,58]	(-0.44, -0.17)
ALogP	Constitutional CDK descriptor	ALogP value (Ghose-Crippen logK_{ow}) [7,59]	(-0.79, 6.44)
MDEC.13	Topological CDK descriptors	Distance edge between all primary and tertiary carbons [60]	(0, 10.65)
MDEO.11		Distance edge between all primary oxygens [60]	(0, 2.39)
khs.sCH3		A fragment count descriptor that uses Kier and Hall smarts fragments (details in Supporting Information Fig. S4) [61]	(0, 7)
khs.dCH2			(0, 2)
khs.sssCH			(0, 4)
khs.ssS			(0, 1)
VC.5		The valence Kier-Hall Chi cluster descriptor of order 5 (details in Supporting Information Fig. S5) [61]	(0, 0.52)

The MLR model was built up using the 11 selected descriptors as given in Eq. 1,

$$logP_{tol/w} = \sum_{i=1}^{n} c_i d_i + b \quad (1)$$

where n stands for the selected descriptors (d_i), c_i denotes the coefficient of descriptor i, and b is the intercept of the regression model.

The RFR model (Eq. 2) relies on randomly created decision trees, which in turn were based on bootstrap aggregating for creating the bootstrapped dataset for the regression according to the

243 *randomForest* library of R [56]. The optimal number of variables tried at each split was 3 and the
244 forecasted value is given as an averaged output across all trees obtaining a mean of squared
245 residuals of 0.97 log*P* units. The importance of each variable for this model is shown in
246 Supporting Information Fig. S2.

$$logP_{tol/w} = \frac{1}{T}\sum_{i=1}^{n} t_i(X_E) \quad (2)$$

247 where (X_E) stands for the ensemble prediction (11 descriptors) and t_i is an average of individual
248 decision trees over the total number of trees (T = 187 in our model; see Supporting Information
249 Fig. S2).
250 Finally, validation of the ML models was carried out through the *k*-fold cross-validation (*k = 5;*
251 see Supporting Information Table S2).
252
253 **Parametrization of the IEF-PCM/MST Solvation Model for Toluene and Benzene.**
254 The IEF-PCM/MST solvation model has been parametrized at the B3LYP/6-31G(d) level to
255 estimate the solvation free energy of organic compounds in a variety of solvents, which include
256 water [62], *n*-octanol [63], chloroform [64], carbon tetrachloride [65], esters and ketones
257 (manuscript in preparation). Following these studies, the aim of this work is to extend the IEF-
258 PCM/MST formalism to the solvation in benzene and toluene. To this end, a training set of
259 experimental solvation free energy values (ΔG_{sol}) in toluene and benzene (91 and 72 compounds,
260 respectively; see Supporting Information qm_sets.xlsx) was compiled. The experimental
261 ΔG_{sol} values were taken from either the Minnesota database [66] or by combining the
262 experimental partition coefficient between the organic solvent and water collected in the
263 literature [24,28,50,67,68] with the hydration free energy of the compound according to Eq. 3.

$$\Delta G_{sol}^{org} = \Delta G_{sol}^{w} - 2.303 \cdot R \cdot T \cdot logP_{org/w} \quad (3)$$

where ΔG_{sol}^{org} and ΔG_{sol}^{w} are the experimental solvation free energies for the organic (benzene, toluene) solvent and water, respectively. At this point, let us note that the ΔG_{sol}^{w} values were taken from the Minnesota database.

In the IEF-PCM/MST model ΔG_{sol} is calculated by combining electrostatic (ΔG_{ele}) and non-electrostatic ($\Delta G_{n\text{-}ele}$) terms (Eq. 4), which are computed using a double molecule-shaped cavity for the solute embedded in the polarizable continuum medium [33,46].

$$\Delta G_{sol} = \Delta G_{ele} + \Delta G_{n-ele} \quad (4)$$

ΔG_{ele} is determined using a solvent-exposed surface generated by scaling the atomic radii by a solvent-dependent factor, λ, which adopted values of 1.25, 1.36, 1.50, 1.54, 1.60, and 1.80 for the solvation of neutral compounds in water, ketones (methyl ethyl ketone, cyclohexanone, and 4-methyl-2-pentanone), *n*-octanol, esters (ethyl and butyl acetate), chloroform, and carbon tetrachloride. Here, the values of λ adopted for toluene and benzene were derived through interpolation of the curve obtained for the representation of the scaling factor versus the dielectric constant for the parametrized solvents (see Supporting Information Fig. S6). For ionic species, the λ values assigned to atoms bearing a formal charge in charged compounds are further reduced by 10% [62].

$\Delta G_{n\text{-}ele}$ is obtained by adding cavitation (ΔG_{cav}) and van der Waals (ΔG_{vW}) contributions, which are determined using the solute's van der Waals surface. ΔG_{cav} is determined by weighting the contribution of the isolated atom determined using Pierotti's formalism, $\Delta G_{P,i}$ (Eq. 5), by the ratio between the solvent-exposed surface of the atom (S_i) and the total surface (S_T).

$$\Delta G_{\text{cav}} = \sum_{i=1}^{N} \frac{S_i}{S_{\text{T}}} \Delta G_{\text{P},i} \quad (5)$$

Finally, ΔG_{vW} is determined using a linear relationship with the surface of each atom (Eq. 6),

$$\Delta G_{\text{vdw}} = \sum_{i=1}^{N} \xi_i S_i \quad (6)$$

where the atomic surface tensions (ξ_i) are determined by fitting the residual term ($\Delta G_{\text{sol}}^{\text{res}}$) obtained by subtracting both electrostatic and cavitation contributions to the experimental solvation free energy (Eq. 7).

$$\Delta G_{\text{sol}}^{\text{res}} = \Delta G_{\text{sol}}^{\text{exp}} - \Delta G_{\text{ele}} - \Delta G_{\text{cav}} \quad (7)$$

Taking advantage of the 252 experimental log$P_{\text{tol/w}}$ values compiled for developing the ML models, 125 compounds (see Supporting Information qm_sets.xlsx) were randomly selected as an external test set for the parametrization of the IEF-PCM/MST model in toluene. In the same way, 40 experimental benzene/water partition coefficient values (see Supporting Information qm_sets.xlsx) were obtained from the literature [67,68] to define an external test set for benzene. Open Babel 3.0.1 was used to explore the conformational flexibility of the compounds with rotatable bonds to generate a set of initial conformations for subsequent refinement calculations. This was accomplished using the Open Babel genetic algorithm, which is a stochastic generator that generates distinct conformers based on the estimated energy and geometrical distance between selected conformations. All geometries were submitted to different QM calculations using Gaussian 16 [69]. The initial set of conformations was optimized at the B3LYP/6-31G(d) level of theory taking into account solvation effects in water, benzene, and toluene using the parametrized IEF-PCM/MST method. The minimum energy nature of all geometries was verified upon inspection of the vibrational frequencies determined using the harmonic oscillator-

rigid rotor model, eliminating those structures displaying negative frequencies and redundant conformers generated along the QM geometry optimization. Furthermore, those conformers that would account for a population < 3% were discarded. Then, a combination of single-point calculations in the gas phase and solvation free energies determined at the IEF-PCM/MST B3LYP/6-31G(d) level for the final conformers was corrected by adding the thermal and entropic corrections. Finally, the partition coefficient was estimated from the Boltzmann-weighted population of the conformational space sampled in both aqueous and organic solvents.

Results and discussion

Experimental determination of the toluene/water partition coefficient in SAMPL9

The log$P_{tol/w}$ values of the 16 compounds selected for the SAMPL9 blind challenge (Fig. 1) were determined by using the potentiometric technique, which requires the use of different partition ratios of toluene and water volumes (Supporting Information Table S1). The recommended volumes were estimated according to the log$P_{tol/w}$ values calculated using Abraham's LSER model [25], which is one of the most useful approaches for the analysis and prediction of partition free energies in chemical and biological systems. The LSER model relies on five solute descriptors (*E*, *S*, *A*, *B* and *V*; Eq. 8), which encode information about the potential interactions formed by the solute in a given solvent.

$$logP = c + e \cdot E + s \cdot S + a \cdot A + b \cdot B + v \cdot V \quad (8)$$

where *E* and *S* refer to the excess molar refraction and dipolarity/polarizability descriptors of the solute, respectively, *A* and *B* are measures of the solute hydrogen-bond acidity and basicity, respectively, and *V* corresponds to the molar volume of the solute.

The coefficients and constants (*c, e, s, a, b*, and *v*) for toluene as a partition solvent were taken from the data reported by Abraham *et al*. [25] (Eq. 9), and the solute descriptors necessary for the determination of log$P_{tol/w}$ for the SAMPL9 set of compounds were determined by using the UFZ-LSER database [31] (see Table 2).

$$logP_{tol/w} = 0.143 + 0.527 \cdot E - 0.720 \cdot S - 3.010 \cdot A - 4.824 \cdot B + 4.545 \cdot V \quad (9)$$

Table 2 Solute descriptor parameters and estimated log$P_{tol/w}$ values of the 16 compounds used in the SAMPL9 blind challenge by using Eq. 9.

Substance	*E*	*S*	*A*	*B*	*V*	log$P_{tol/w}$ (Eq.9)
Albendazole	1.974	1.96	0.65	1.08	1.948	1.46
Alprenolol	1.250	1.09	0.15	1.44	2.159	2.43
Amitriptyline	1.920	1.42	0.00	1.08	2.400	5.83
Bifonazole	2.410	2.13	0.00	1.15	2.501	5.70
Chlorpheniramine	1.465	1.41	0.00	1.33	2.210	3.53
Epinephrine	1.400	1.17	1.21	1.53	1.415	-4.55
Fluphenazine	2.160	2.25	0.18	1.85	3.091	4.24
Glyburide	2.740	3.85	0.85	2.01	3.558	2.73
Imipramine	1.150	1.45	0.00	1.04	2.402	5.60
Ketoprofen	1.650	2.26	0.55	0.89	1.978	2.43
Nalidixic acid	1.630	1.86	0.01	1.16	1.700	1.76
Paracetamol	1.120	1.63	1.01	0.91	1.172	-2.54
Pindolol	1.700	1.59	0.30	1.40	2.009	1.37
Quinine	2.469	1.23	0.37	1.97	2.551	1.54
Sulfamethazine	2.130	2.53	0.59	1.53	2.004	-0.60
Trazodone	2.800	2.44	0.00	1.68	2.730	4.17

Table 3 reports the measured log$P_{tol/w}$ values determined from at least three independent titrations in conjunction with the experimental equipment used for each compound in the dataset. The log$P_{tol/w}$ values span a range of 7.1 log$P$ units. The most hydrophilic compound is paracetamol (log$P_{tol/w}$ = -1.59), whereas the largest solubility in toluene is found for amitriptyline (log$P_{tol/w}$ = 5.51) and bifonazole (log$P_{tol/w}$ = 5.47). With the exception of paracetamol, epinephrine and sulfamethazine, which exhibit negative log$P_{tol/w}$ values, all the compounds show a preferential partition in toluene. It is worth noting that the standard deviation is generally lower than 0.07 (and in most cases lower than 0.03), but for epinephrine, where the standard deviation is 0.14. The experimental log$P_{tol/w}$ values determined for the SAMPL9 challenge were added to our database (entries 253-268, see Supporting Information DatabaseTol_feb23.csv).

Table 3 Experimental log$P_{tol/w}$ values determined for the SAMPL9 dataset.

ID	Name	log$P_{tol/w}$	Standard deviation	Exp. equipment
1	Albendazole	3.76	0.03	Pion SiriusT3
2	Alprenolol	2.40	0.01	Pion SiriusT3
3	Amitriptyline	5.51	0.01	Sirius D-PAS & GLpKa
4	Bifonazole	5.47	0.06	Pion SiriusT3
5	Chlorpheniramine maleate salt	3.61	0.01	Sirius D-PAS & GLpKa
6	Epinephrine	-1.23	0.14	Pion SiriusT3
7	Fluphenazine dihydrochloride	4.37	0.04	Sirius D-PAS & GLpKa
8	Glyburide	2.79	0.02	Pion SiriusT3
9	Imipramine hydrochloride	5.05	0.02	Sirius D-PAS & GLpKa
10	Ketoprofen	2.47	0.02	Pion SiriusT3
11	Nalidixic acid	1.46	0.01	Sirius D-PAS & GLpKa
12	Paracetamol	-1.59	0.06	Pion SiriusT3
13	Pindolol	0.36	0.01	Pion SiriusT3
14	Quinine	1.41	0.01	Sirius D-PAS & GLpKa
15	Sulfamethazine	-0.74	0.03	Sirius D-PAS & GLpKa
16	Trazodone hydrochloride	3.77	0.07	Pion SiriusT3

Prediction of log$P_{tol/w}$ by MLR and RFR models

A variety of molecular descriptors have been exploited in MLR-based studies aiming to estimate the partition coefficient, such as parameters describing the features of the chemical groups (or the whole molecule) present in the solute [70], holograms built for each compound, where the hologram consists of the count of each atom-type [71], surface area and volume [72, 73], hydrophobic area and chain descriptors [73], and QM-based electronic properties [74]. On the other hand, more complex ML methods such as RFR have been used to describe the quantitative structure-property relationships with lipophilicity [75, 76]. Following these studies, MLR and

RFR are used here for predicting the log$P_{tol/w}$ values of 252 neutral compounds that encompass a diverse range of physicochemical properties (Fig. 3).

The MLR model developed in this work relies on the final set of 11 relevant descriptors described in Table 1. The determination coefficient (r^2) between the descriptors and the log$P_{tol/w}$ values for the 252 compounds included in the dataset is <0.11. The only exception is AlogP (i.e., the calculated log$P_{o/w}$ determined using Ghose-Crippen atomic contributions [7,59]), where the determination coefficient is 0.58.

The MLR model was fitted following Eq. 1 for a subset of 214 compounds included in the training set (r^2 = 0.75, s = 1.00, F = 54.97, p-value = 2.2 x 10^{-16}), and the weight of each descriptor is represented in Fig. 5. The presence of fragments such as methyl, exo-methylene, aliphatic methine, and sulfide groups, and double bond conjugated to a carbonyl group favor the solvation in toluene. On the other hand, the occurrence of higher count of halogens, distances between primary carbons and oxygens, and tertiary carbons, charge weighted descriptors, and Chi clusters descriptors of order 5 (found in catechol, anisole, and indole fragments) favors the partition in the aqueous phase. In addition, depending on its sign, ALogP modulates at large extent the preference of a compound for toluene or water, which is in agreement with the trends reported in previous studies [30,77,78].

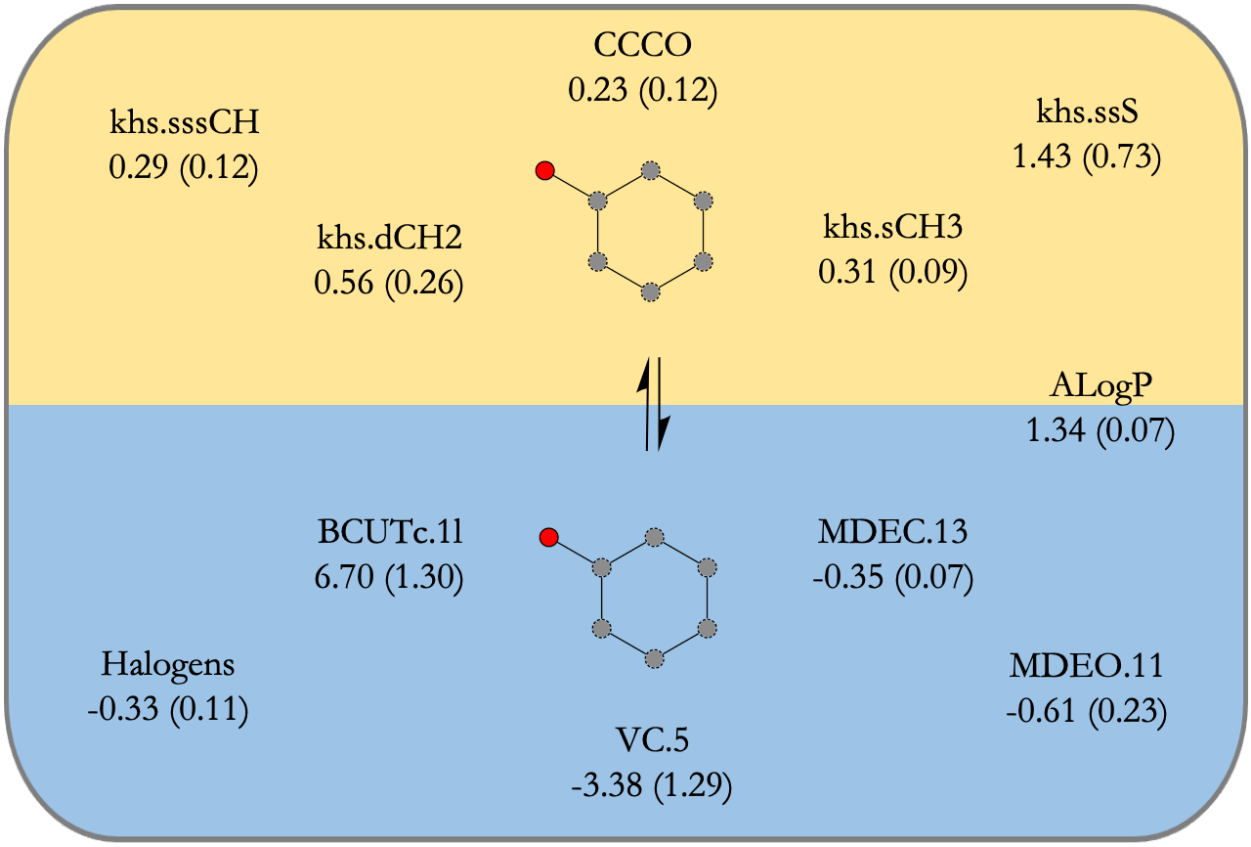


Fig. 5 Graphical representation of the descriptors that favor the partition of a given compound in toluene (yellow) and water (blue) according to the MLR model fitted following Eq. 1. The weight of each descriptor is given below the label (the standard deviation is indicated in parenthesis). The intercept of the model is 1.78 (0.55).

The RFR approach was used as an alternative method to develop a predictive model. For the sake of simplicity, the RFR model used the default number of variables to be considered in each split (11 features / 3 ~ 3) in the randomForest package. On the other hand, the number of trees with the lowest mean square error (MSE) was 187, providing an average mean squared error of 0.97 logP units (see Supporting Information Fig. S2). ALogP and at less extent BCUTc.1l are found to be the most relevant properties in the fitting to the training set, whereas a similar contribution is found for the rest of descriptors, as deduced from the analysis of the mean decrease accuracy (Supporting Information Fig. S2).

Comparison of the results obtained for MLR and RFR models for both training and test sets is shown in Fig. 6 (see also Table 4). The RFR model was slightly superior to the MLR model for the fitting of the training set in all metrics (RMSE of 0.5 and 1.0 log*P* units), but the two models exhibit a similar accuracy for the compounds in the test set (RMSE of 1.1 log*P* units for both MLR and RFR). In addition, the results of the 5-fold cross-validation were also highly similar for the two models, as noted in RMSE values close to 1.0 log*P* units (Supporting Information Table S2).

Table 4 Statistical parameters of the comparison between experimental and predicted $logP_{tol/w}$ values for the training and test set using the regression models.

Method	r^2		RMSE		MUE		MSE	
	Training	Test	Training	Test	Training	Test	Training	Test
MLR	0.75	0.58	0.97	1.13	0.72	0.89	0.00	-0.16
RFR	0.95	0.67	0.51	1.05	0.38	0.80	0.00	-0.14

r^2: determination coefficient; RMSE: root-mean-square error; MUE: mean unsigned error; MSE: mean signed error (log*P* units.)

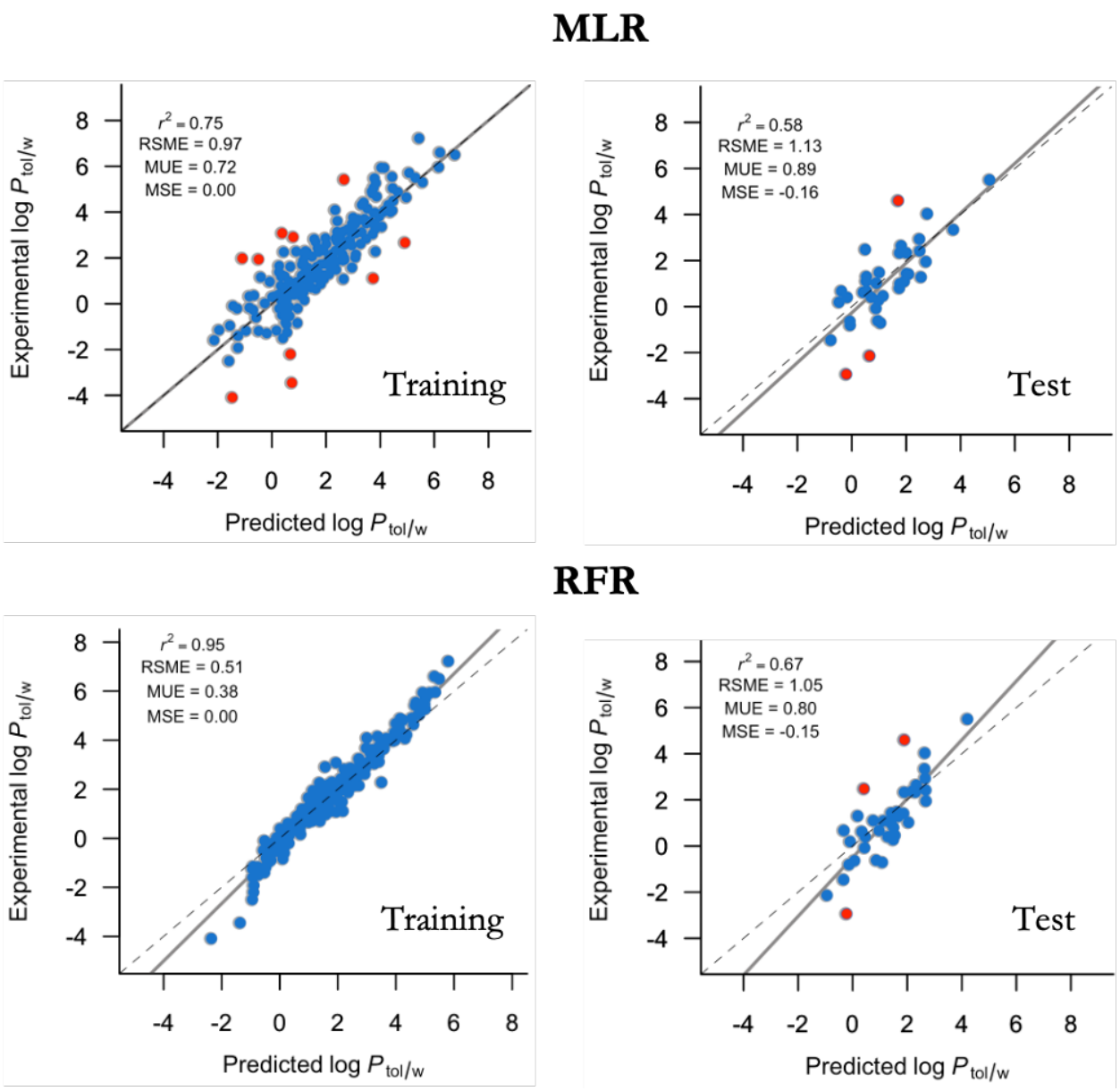


Fig. 6 Comparison between experimental and calculated log$P_{tol/w}$ values. Results obtained using (top) the MLR and (bottom) RFR methods for the (left) training and (right) test sets. Red dots highlight compounds with a deviation greater than 2 log*P* units (MLR - Training: hydroquinone, gallic acid, thymine, diantipyrylmethane, diantipyrylphenylmethane, 2,3,4-trimethylpentane, 2,4-dinitrophenol, benzoylacetone, N-methylthalidomide, N-propylthalidomide; MLR – Test: resorcinol, pyrogallol, N-pentylpthalidomide; RFR – Test: 2'-hydroxyacetophenone, pyrogallol, N-eentylpthalidomide; see Supporting Information DatabaseTol_feb23.csv).

Extension of the IEF-PCM/MST model to toluene and benzene

Within the framework of QM methods, the QM-SCRF formalism provides an effective way to examine the effect of solvation on the solute's properties and to explore solvent-induced changes in chemical processes, such as the thermodynamics and kinetics of reactions and spectral properties of compounds [32]. The success of these methods relies on the development of parametrized QM-SCRF methods and the computational efficiency provided by the implicit treatment of solvent molecules [79,80]. In this context, extension of the IEF-PCM/MST method to toluene would expand the range of (bio)chemical applications of this continuum solvation model.

Inspection of the set of 91 solvation free energies in toluene revealed the lack of compounds containing P and S atoms, which would thus affect the parametrization of the surface tensions used in the IEF-PCM/MST model for these atoms (Eq. 6). To circumvent this limitation, solvation free energies in benzene were compiled for 72 compounds, including compounds containing P and S, although this set lacked F-containing compounds. At this point, it is worth noting that 52 solutes common to both toluene and benzene exhibit almost identical solvation free energies (spanning values from -1.8 to -10.0 kcal mol^{-1}) in the two aromatic solvents (Fig. 7), as noted in the regression equation $y = 1.007\ x + 0.14$ ($r^2 = 0.99$; x and y denote the solvation free energies in benzene and toluene, respectively). Therefore, taking into account the similar physicochemical properties of the two solvents (see Supporting Information Table S3), a preliminary adjustment of the surface tensions was performed for the whole set of 163 data reported for the two aromatic solvents. Next, the surface tensions of F, P and S were fixed and the surface tensions of the remaining atom types were optimized by adjusting separately the experimental data in either toluene or benzene.

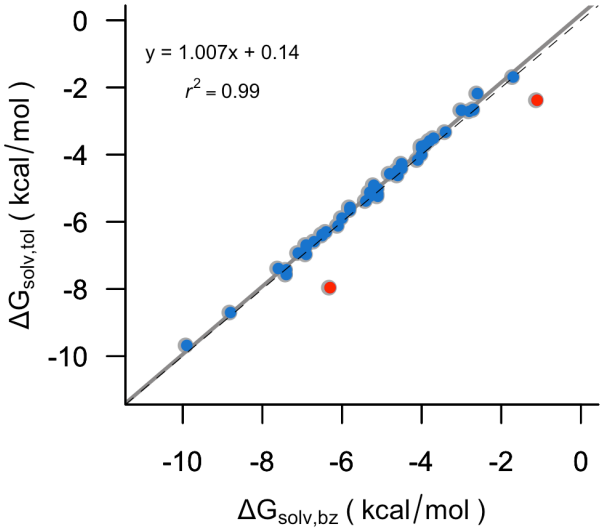


Fig. 7 Comparison of the experimental solvation free energies (kcal mol^{-1}) measured for a subset of 50 solutes common to toluene (tol) and benzene (bz). Red dots highlight compounds with a deviation greater than 1 kcal/mol units, which correspond to ammonia (ΔG_{sol} of -1.1 and -2.4 kcal mol^{-1} in benzene and toluene, respectively) and methyl benzoate (ΔG_{sol} of -6.3 and -8.0 kcal mol^{-1} in benzene and toluene, respectively).

Table 5 gives the atomic surface tensions optimized for the IEF-PCM/MST method parametrized at the B3LYP/6-31G(d) level for toluene and benzene. As expected from the preceding analysis, the atomic surface tensions determined in toluene and benzene are highly similar, and the slight differences observed for a given atom type in the two solvents are caused by the small discrepancies found in the calculation of the cavitation free energy.

The comparison between experimental and fitted ΔG_{sol} values in toluene and benzene is shown in Fig. 8. In both solvents, the RMSE and MUE are close to 0.8 and 0.6 kcal mol^{-1}, respectively. Thus, the parametrized IEF-PCM/MST model provides results similar to those obtained using other continuum solvation methods for the solvation of neutral compounds in non-aqueous solvents (e.g., COSMOtherm, SM8 and IEF-PCM/MST with MUE close to 0.6 kcal mol^{-1} [80]).

Table 5 Optimized atomic surface tensions (ξ_i; kcal $mol^{-1}Å^{-2}$) determined for IEF-PCM/MST calculations at the B3LYP/6-31G(d) level for the whole training data set or separately for toluene and benzene.

Atom type	Whole dataset	Toluene	Benzene
C	-0.199	-0.194	-0.204
CH	-0.206	-0.198	-0.215
CH_2	-0.202	-0.194	-0.209
CH_3	-0.193	-0.189	-0.207
N	-0.262	-0.257	-0.279
NH	-0.258	-0.249	-0.262
O	-0.235	-0.223	-0.214
OH	-0.260	-0.237	-0.265
F	-0.144	-0.144	-0.144
P	-0.183	-0.183	-0.183
S	-0.183	-0.183	-0.183
SH	-0.183	-0.183	-0.183
Cl	-0.184	-0.178	-0.198
Br	-0.193	-0.197	-0.222

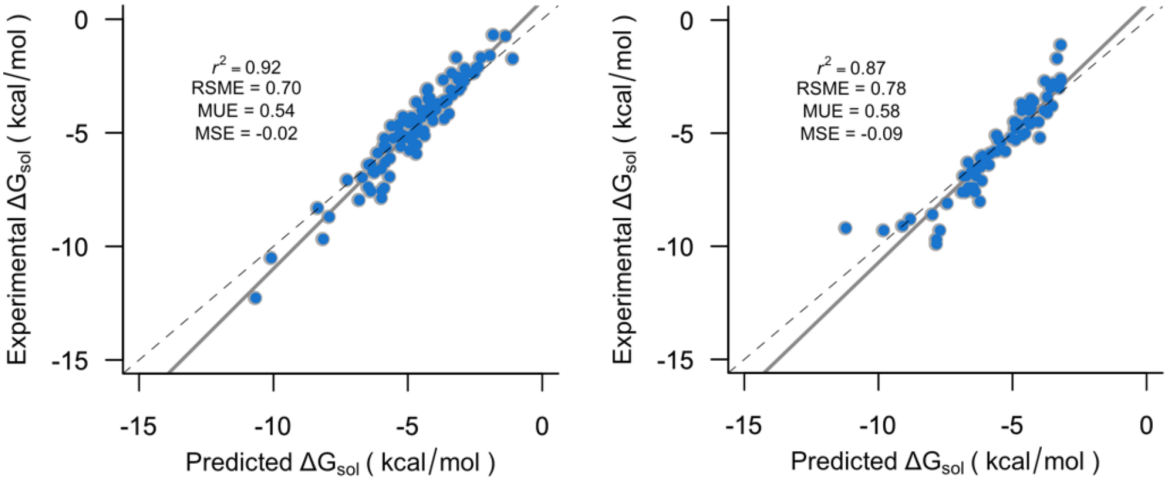


Fig. 8 Comparison between experimental and predicted solvation free energies in (left) toluene and (right) benzene for the compounds used in the parametrization of the IEF-PCM/MST model.

Finally, to further calibrate the parametrization of the IEF-PCM/MST model, additional calculations were performed to estimate toluene/water and benzene/water partition coefficients for an external test set of 125 and 40 small compounds, respectively (Fig. 9). The results pointed out that reliable estimates of the partition coefficient are obtained in the two solvents (RSME < 1 log*P* units), which agree with those previously reported for the IEF-PCM/MST method in the *n*-octanol/water system applied to nitrogen-containing aromatic compounds (RSME = 0.8 log*P* units), drug-like compounds (RSME = 1.1 log*P* units) [81], fragment-like small molecules in the SAMPL6 log*P* challenge (RSME = 0.8 log*P* units) [82], and sulfonamide-containing compounds in the SAMPL7 log*P* contest (RSME = 1.0 log*P* units) [83]. The similar performance for the solvation in distinct solvents is likely determined by the similar number of experimental data used in the parameterization of the model (91, 72, and 58 for toluene, benzene, and *n*-octanol [63], respectively). Overall, these results support the consistency of the parametrization

performed for toluene and benzene within the current formalism of the IEF-PCM/MST model for the prediction of both solvation free energies as well as partition coefficients.

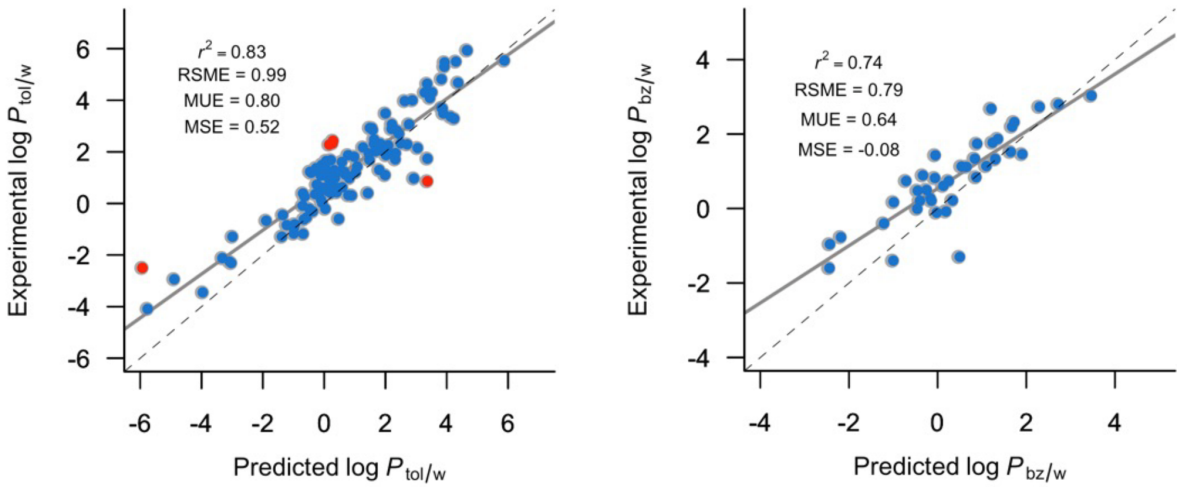


Fig. 9 Comparison between experimental and predicted (left) toluene/water and (right) benzene/water partition coefficients for the compounds used as a test set in the parametrization of IEF-PCM/MST model. Red dots highlight compounds with a deviation greater than 2 log*P* units (4-(trifluoromethyl)aniline, sulfanilamide, 1-naphthylamine, 2-naphthylamine, and nicotine)

Prediction of log$P_{tol/w}$ for the SAMPL9 dataset

The external set of 16 compounds in SAMPL9 was used to further check the performance of the ML and IEF-PCM/MST models discussed above (Table 6). A conformational sampling with Open Babel was performed to explore the conformational preferences for each compound, as described in Materials and Methods. The final set of conformers was visually checked, leading between 2 and 10 conformers for the SAMPL9 test compounds. Abraham´s LSER equation was also included in the comparison for the sake of comparison.

519 The RMSE between experimental data and the log$P_{tol/w}$ values estimated with the LSER equation
520 is 1.1 log$P$ units and the determination coefficient is close to 0.9, which points out the agreement
521 between predicted and experimental log$P_{tol/w}$ values for most of the compounds. Indeed, only
522 two compounds (albendazole and epinephrine) showed a deviation larger than 2 log$P$ units
523 toward partition in the aqueous phase relative to the measured values (Fig. 10). Without
524 detracting from the good performance of Abraham's equation, there is evidence of the formation
525 of intramolecular hydrogen bonds in both albendazole [84] and epinephrine [85,86], which
526 would thus favor an increase in the lipophilicity of these compounds [87]. Exclusion of these
527 compounds reduces the RMSE to 0.4 log$P$ units, with a mean signed error of only 0.1 log$P$ units,
528 and a determination coefficient of 0.97.
529
530
531
532
533
534
535
536
537
538
539
540
541

Table 6 Experimental and calculated partition coefficients (log$P_{tol/w}$) for the 16 compounds included in the SAMPL9 blind challenge.[a]

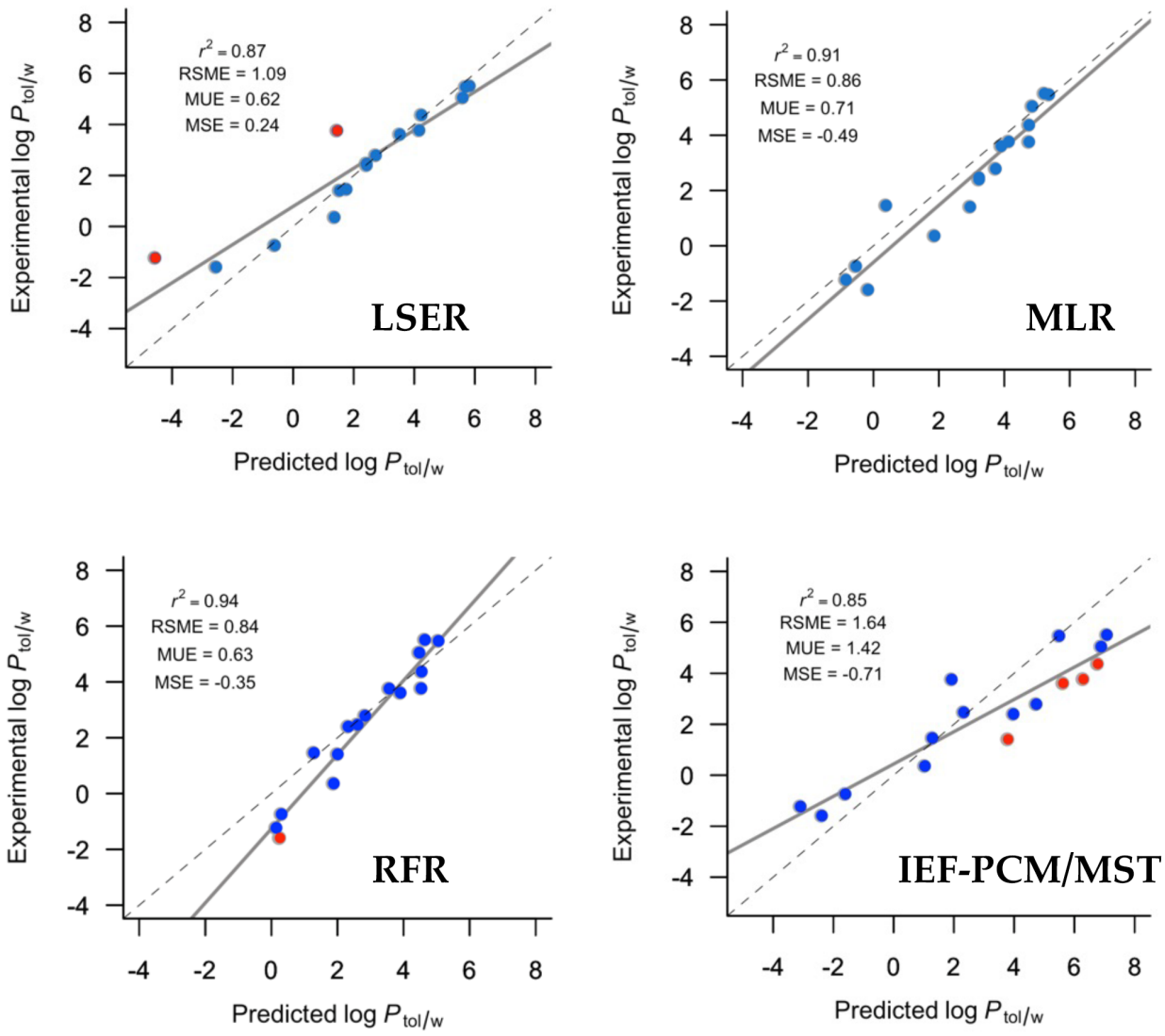
Compound	LSER	MLR	RFR	IEF-PCM/MST [b]	Exptl.
Albendazole	1.46	4.76	3.57	1.93 (1.94)	3.76
Alprenolol	2.43	3.23	2.33	3.98 (4.71)	2.40
Amitriptyline	5.83	5.23	4.65	7.08 (6.98)	5.51
Bifonazole	5.70	5.37	5.06	5.50 (5.70)	5.47
Chlorpheniramine	3.53	3.91	3.91	5.63 (5.44)	3.61
Epinephrine	-4.55	-0.84	0.15	-3.09 (-3.19)	-1.23
Fluphenazine	4.24	4.76	4.55	6.78 (6.97)	4.37
Glyburide	2.73	3.74	2.84	4.74 (5.40)	2.79
Imipramine	5.60	4.86	4.48	6.89 (6.79)	5.05
Ketoprofen	2.43	3.23	2.61	2.33 (2.37)	2.47
Nalidixic acid	1.76	0.39	1.29	1.29 (1.26)	1.46
Paracetamol	-2.54	-0.16	0.26	-2.38 (-2.44)	-1.59
Pindolol	1.37	1.87	1.89	1.04 (0.92)	0.36
Quinine	1.54	2.96	2.01	3.80 (3.60)	1.41
Sulfamethazine	-0.60	-0.53	0.31	-1.60 (-1.73)	-0.74
Trazodone	4.17	4.14	4.54	6.30 (6.86)	3.77
r^2	***0.87***	***0.91***	***0.94***	***0.85 (0.84)***	
RMSE	***1.09***	***0.86***	***0.84***	***1.64 (1.72)***	
MUE	***0.62***	***0.71***	***0.63***	***1.42 (1.51)***	
MSE	***0.24***	***-0.49***	***-0.35***	***-0.71 (-0.8)***	

[a] r^2: determination coefficient; RMSE, root-mean-square error; MUE: mean unsigned error; MSE: mean signed error in log P units.

[b] Data obtained by combining gas phase energies determined at the MP2/aug-cc-pVDZ with the solvation free energy estimated from IEF-PCM/MST B3LYP/6-31G(d) calculations are given in parenthesis.

The MLR and RFR models yielded log$P_{tol/w}$ values in close agreement with the experimental data, as noted in determination coefficients comprised between 0.91 and 0.94 and errors slightly lower than 0.9 log$P$ units. No compound was predicted to have a deviation larger than 2 log$P$ units with the MLR model, and only paracetamol was predicted to be too lipophilic with the RFR model (Fig. 10). A similar, but less pronounced effect was found for epinephrine, which was also predicted to partition better in toluene, in contrast with the hydrophilic character observed in the

556 experimental value. Overall, these results are encouraging taking into account that in the
557 SAMPL5 challenge [78], aimed at the blind prediction of cyclohexane/water distribution
558 coefficients for a set of 53 small molecules, just two submissions used empirically trained
559 methods obtaining an RMSE of 3.0 [79] and 3.3 [80], COSMO-RS being the best ranked method
560 with an RMSE of 2.1 log*P* units [81].
561
562
563
564
565



566

567 **Fig. 10** Comparison between experimental and predicted log$P_{tol/w}$ for the 16 compounds used as
568 an external set in this work with (top left) Abraham´s Equation, (top right) MLR, (bottom left)
569 RFR and (bottom right) IEF-PCM/MST methods. Red dots highlight compounds with a
570 deviation greater than 2 log$P$ units (LSER: albendazole, epinephrine; RFR: paracetamol; IEF-
571 PCM/MST: chlorpheniramine, fluphenazine, quinine, trazodone).
572

573 Compared to the previous approaches, the IEF-PCM/MST method exhibits a lower performance
574 for the prediction of the log$P_{tol/w}$ values (Fig. 10). Although the determination coefficient
575 compares with the value obtained for the LSER model, the error amounts to 1.6 log$P$ units. This
576 can be attributed to the overestimated lipophilicity of chlorpheniramine, fluphenazine, quinine

577 and trazodone. All these compounds have high flexibility (> 4 rotatable bonds) and contain, with
578 the exception of chlorpheniramine, four cyclic systems. In order to examine the potential effect
579 due to an inaccurate description of the balance between the internal stability between conformers
580 and the solvation in toluene and water, additional single-point calculations were performed at the
581 MP2/aug-cc-pVDZ level to estimate the intramolecular (gas phase) energy of conformers, which
582 was subsequently combined with the solvation free energy in the organic and aqueous phases.
583 The computed $\log P_{\mathrm{tol/w}}$ values are very similar to the results obtained directly from IEF-
584 PCM/MST B3LYP/6-31G(d) calculations, as can be noted in Table 10, thus suggesting that the
585 deviation from experimental data obeys to other factors.
586 Clearly, the LSER equation and the ML-derived models benefit from the direct fitting of the
587 $\log P_{\mathrm{tol/w}}$ data, which is the target property used in the parametrization of the weighting factors
588 that modulate the contribution of the molecular descriptors. In contrast, the target property in the
589 IEF-PCM/MST model is the solvation free energy, and therefore the error in the estimated
590 $\log P_{\mathrm{tol/w}}$ is affected by the addition of the uncertainties in the computation of ΔG_{sol} in toluene and
591 water. Nevertheless, this strategy broadens the range of chemical processes that may be
592 examined for solutes in both organic and aqueous phases with the IEF-PCM/MST model.
593 On the other hand, it is also clear that the limited number of experimental data available for the
594 solvation in toluene challenges the statistical significance of the atomic surface tensions fitted for
595 certain types of atoms, which are underrepresented in the training dataset (see above). Here this
596 limitation has been addressed by including data relative to the solvation in benzene, taking
597 advantage of the similar structural and chemical properties of the two aromatic solvents, which is
598 reflected in the close correspondence observed for the solvation free energy in toluene and
599 benzene for the subset of compounds available in the two solvents. While this strategy may be

valuable for the extension to other kind of organic solvents, present results point out the need to consider not only a proper classification of the distinct atom types included in the solvation model, but also to ensure a proper coverage of the distinct atom types in the experimental dataset.

Conclusions

In contemporary drug discovery research, obtaining reliable values of toluene/water partition coefficients is essential to examine relevant features of drugs, such as the propensity to adopt conformations with intramolecular hydrogen bonds, which may have a relevant influence on the ADME-Tox profile. In this context, the experimental results determined for the set of 16 drugs reported here are expected to be valuable for calibrating the accuracy and consistency of current empirical, physical-based and simulation methods developed for the prediction of partition coefficients.

In this study we have developed two machine-learning models based on multiple linear regression (MLR) and random forest regression (RFR), which exhibit an excellent performance in the prediction of the $\log P_{tol/w}$ values. The models exhibit a slightly better performance compared to Abraham's LSER model, which can be attributed to the selection of the descriptors encoded in the two models. Given its easy implementation and low computational cost, they constitute an excellent option to obtain reliable data for drug-like compounds.

On the other hand, the parametrization of the B3LYP/ 6-31G(d) version of the integral IEF-PCM/MST continuum solvation model has addressed the solvation of small compounds in benzene and toluene as a strategy to override the scarce representation of data for certain types of atoms. While the results support the suitability of this approach, the prediction of $\log P_{tol/w}$ values

623 is affected by the uncertainties associated to the solvation free energy in the two solvents,
624 although this also confers larger flexibility to tackle the study of a wider range of chemical
625 processes in these solvents and gain insight into the solvent-induced changes in electronic
626 properties and chemical reactivity.
627 Overall, bearing in mind the relatively limited data available for solvent systems beyond *n*-
628 octanol/water, the results are encouraging and support the use of diverse computational strategies
629 to estimate the solvation free energies and partition coefficients of (bio)organic compounds in
630 drug discovery.
631

632 ## Conflicts of interest

633 There are no conflicts to declare.
634

635 ## Acknowledgments

636 The authors thank the Vice Chancellor for Research of the University of Costa Rica for its
637 support work via the research projects 115-C2-126 and 908-C3-610. C. C., C. R. and F. J. L
638 acknowledge financial support from the State Research Agency/Spanish Ministry of Science and
639 Innovation (AEI/10.13039/501100011033; grants MDM-2017-0767, PID2020-115812GB-I00,
640 PID2020-115374GB-100, PID2020-117646RB-I00 and CEX2021-001202-M) and the
641 Generalitat de Catalunya (2021SGR00671). The Consorci de Serveis Universitaris de Catalunya
642 (CSUC) is acknowledged for providing computational resources (Molecular Recognition
643 project). We are grateful to Prof. D. L. Mobley (UC Irvine) for the coordination of the SAMPL9
644 blind challenge with the support of the National Institutes of Health (R01GM124270).

References

1. B. Testa, P. Crivori, M. Reist and P.-A. Carrupt, *Perspect. Drug Discov. Des*. 2000, **19**, 179–211.
2. J. A. Arnott and S. L. Planey, *Expert Opin. Drug Discov*. 2012, **7**, 863–875.
3. T. W. Johnson, R. A. Gallego and M. P. Edwards, *J. Med. Chem*. 2018, **15**, 6401–6420.
4. D. A. DeGeoy, H. J. Chen, P. B. Cox and M. D. Wendt, *J. Med. Chem*. 2018, **61**, 2636–2651.
5. M. Janicka, M. Sztanke and K. Sztanke, *Molecules*, 2020, **25**, 487.
6. R. Mannhold and K. Dross, *Quant. Struct.-Act Relat*., 1996, **15**, 403–409.
7. A. K. Ghose, V. N. Viswanadhan and J. J. Wendoloski, *J. Phys. Chem. A*, 1998, **102**, 3762–3772.
8. R. Mannhold and H. Van De Waterbeemd, *J. Comput. Aided Mol. Des*., 2001, **15**, 337–354.
9. R. Mannhold, G. I. Poda, C. Ostermann and I. V. Tetko, *J. Pharm. Sci.*, 2009, **98**, 861–893.
10. M. Isik, T. D. Bergazin, T. Fox, A. Rizzi, J. D. Chodera and D. L. Mobley, *J. Comput. Aided Mol. Des*., 2020, **34**, 335–370.
11. A. Leo, C. Hansch and D. Elkins. *Chem. Rev*., 1971, **71**, 525–616.
12. J. Sangster, *J. Phys. Chem. Ref. Data*., 1989, **18**, 1111–1227.
13. A. J. Leo, *Chem Rev*, 1993, **93**, 1281–1306.
14. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Deliv. Rev.* 1997, **23**, 3–25.
15. C. Steinebach, I. Sosič, S. Lindner, A. Bricelj, F. Kohl, Y. L. D. Ng, M. Monschke, K. G. Wagner, J. Krönke and M. Gütschow, *Medchemcomm*, 2019, **10**, 1037–1041.

16. V. G. Klein, C. E. Twonsend, A: Testa, M. Zengerle, C. Maniaci, S. J. Hughes, K.-H. Chan, A. Ciulli and R. S. Lokey, *ACS Med. Chem. Lett*., 2020, **11**, 1732–1738.
17. Y. Atilaw, V. Poongavanam, C. Svensson Nilsson, D. Nguyen, A. Giese, D. Meibom, M. Erdelyi and J. Kihlberg, *ACS Med. Chem. Lett*., 2021, **12**, 107–114.
18. A. Avdeef. *Curr. Topics. Med. Chem*. 2001, **1**, 277–351.
19. R. A. Saunders and J. A. Platts, *New. J. Chem*. 2004, **28**, 166–172.
20. T. Hartmann and J. Schmidtt, *Drug Discov. Today Technol*. 2004, **1**, 431–439.
21. P. W. Kenny, C.A. Montanari and I.M. Prokopczyk, *J. Comput. Aided Mol. Des*., 2013, **27**, 389–402.
22. E. Danelius, V. Poongavanam, S. Peintner, L. H. E. Wieske, M. Erdélyi and J. Kihlberg, *Chem. Eur. J*., 2020, **26**, 5231–5244.
23. G. Ermondi, M. Vallaro, J. Saame, L. Toom, I. Leito, R. Ruiz and G. Caron, *Eur. J. Pharm. Sci*. 2021, *161*, 105802.
24. R. Ruiz, W. J. Zamora, C. Ràfols and E. Bosch, *Eur. J. Pharm. Sci*., 2022, **168**, 106066.
25. M. H. Abraham, R. E. Smith, R. Luchtefeld, A. J. Boorem, R. Luo and W. E. Acree Jr., *J. Pharm. Sci*., 2010, **99**, 1500–1515.
26. L. David, M. Wenlock, P. Barton and A. Ritzén, *ChemMedChem,* 2021, **16**, 2669–2685.
27. G. Caron and G. Ermondi, *J. Med. Chem*., 2005, **48**, 3269–3279.
28. G. Ermondi, A. Visconti, R. Esposito and G. Caron, *Eur. J. Pharm. Sci*, 2014, **53**, 50–54.
29. G. Caron, J. Kihlberg and G. Ermondi, *Med. Res. Rev*, 2019, **39**, 1707–1729.
30. R. Ruiz, W. J. Zamora, C. Ràfols and E. Bosch. *Eur. J. Pharm. Sci*. 2022, **168**, 106066
31. S. Ulrich, T.N. Brown, N. Watanabe, G. Bronner, M. H. Abraham, K.-U. N. E. Goss, 2017, UFZ-LSER database v 3.2 [Internet]

690 32. J. Tomasi, B. Mennuci and R. Cammi, *Chem. Rev*. 2005, **105**, 2999–3094.
691 33. F. J. Luque, C. Curutchet, J. Muñoz-Muriedas, A. Bidon-Chanal, I. Soteras, A. Morreale, M.
692 Orozco. *Phys. Chem. Chem. Phys*. 2003, **5**, 3827–3836.
693 34. R. E. Skyner, J. L. McDonagh, C. R. Groom, T. van Mourik and J. B. O. Mitchell. *Phys.*
694 *Chem. Chem. Phys*. 2015, **17**, 6174–6191.
695 35. J. M. Herbert, *WIREs Comput. Mol. Sci*. 2021, **11**, e1519.
696 36. J. R. Pliego Jr. and J. M. Riveros, *WIREs Comput. Mol. Sci*. 2020, **10**, e1440.
697 37. A. V. Marenich, C. J. Cramer and D. G. Truhlar, J. Phys. Chem. B, 2009, 113, 6378–6396.
698 38. A. Klamt, *WIREs Comput. Mol. Sci*. 2018, **8**, e1338.
699 39. M. T. Geballe, A. G. Skillman, A. Nicholls, J. P. Guthrie and P. J. Taylor, *J. Comput. Aided*
700 *Mol. Des*. 2010, **24**, 259-279.
701 40. A. G. Skillman, *J. Comput. Aided Mol. Des*. 2012, **26**, 473–474.
702 41. D. L. Mobley, K. L.Wymer, N. M. Lim and J. P. Guthrie, *J. Comput. Aided Mol. Des*. 2014,
703 **28**, 135–150.
704 42. C. C. Bannan, K. H. Burley, M. Chiu, M. R. Shirts, M. K. Gilson and D. L. Mobley, *J.*
705 *Comput. Aided Mol. Des*. 2016, **30**, 927–944
706 43. T. D. Bergazin, N. Tielker, Y. Zhang, J. Mao, M. R. Gunner, K. Francisco, C. Ballatore, S.
707 M. Kast and D. L. Mobley, *J. Comput. Aided Mol. Des*. 2021, **35**, 771-802.
708 44. M. N. Bahr, A. Nandkeolyar, J. K. Kenna, L. Da Và, N. nevins, M. Isik, J. D. Chodera and
709 D. L. Mobley, *J. Comput. Aided Mol. Des*. 2021, **35**, 1141-1155.
710 45. https://github.com/samplchallenges/SAMPL9/tree/main/logP
711 46. I. Soteras, C. Curutchet, A. Bidon-Chanal, M. Orozco and F. J. Luque, *J. Mol. Struct.*
712 *THEOCHEM,* 2005, **727**, 29–40.

47. A. Avdeef, *J. Pharm. Sci.* 1993, **82**, 183–190.
48. C. Ràfols, E. Bosch, R. Ruiz, K. J. Box, M. Reis, C. Ventura, S. Santos, M.E. Araujo and F. Martins, *J. Chem. Eng. Data*, 2012, **57**, 330, 338.
49. C. Ràfols, X. Subirats, J. Rubio, M. Rosés and E. Bosch, *Talanta*, 2017, **162**, 293–299.
50. S. Tshepelevitsh, K. Hernits, J. Jenčo, J. M. Hawkins, K. Muteki, P. Solich and I. Leito, *ACS Omega*, 2017, **2**, 7772–7776.
51. N.M. O'Boyle, M. Banck, C.A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminform*., 2011, **3**, 33.
52. 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.
53. R. Todeschini and V. Consonni, Handbook of Molecular Descriptors, in Methods and Principles In Medicinal Chemistry, ed. R. Mannhold, K. Kubinyi and H. Timmerman, Vol. 11, Wiley-VCH, Weinheim, 2000.
54. https://smarts.plus/
55. https://cran.r-project.org/web/packages/leaps/leaps.pdf
56. L. Breiman, Random Forests. *Mach. Learn*., 2001, **45**, 542–555.
57. R. S. Pearlman and K. M. Smith, *Perspect. Drug Discov. Des*., 1998, **9**, 339–353.
58. R. S. Pearlman and K.M. Smith, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 28–35.
59. A. K. Ghose and G. M. Crippen, *J. Chem. Inf. Comput. Sci*., 1987, **27**, 21–35.
60. S. Liu, C. Cao and Z. Li, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 387–394.

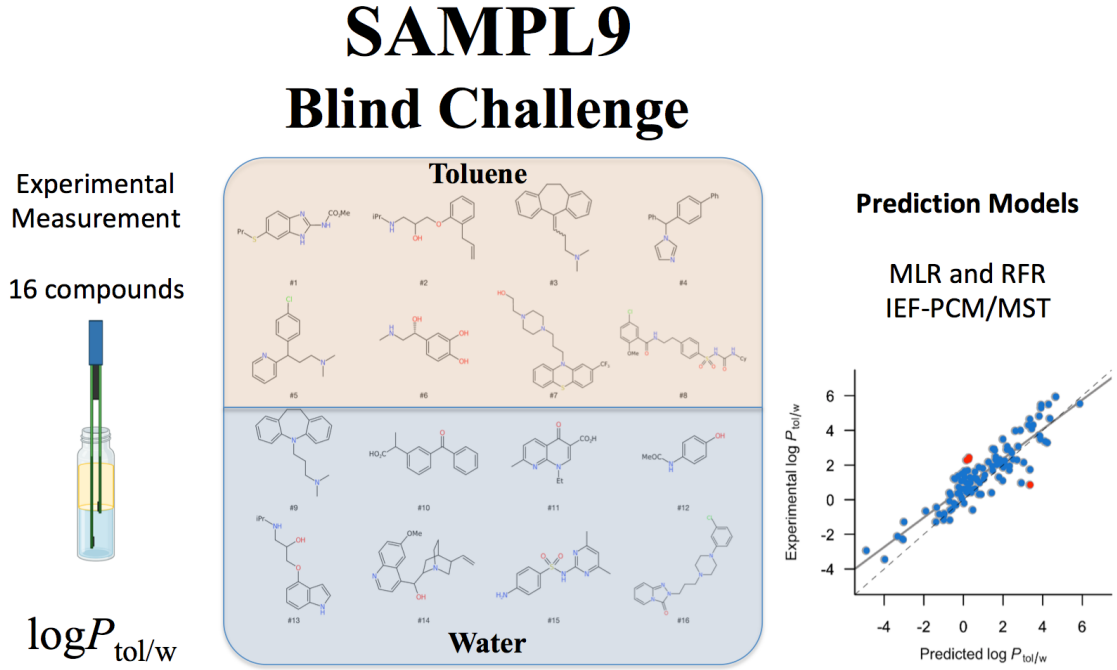
61. L. H. Hall and L. B. Kier, The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling, in *Reviews in Computational Chemistry*, ed. K. B. Lipkowitz and D. B. Boyd, Vol. 2, Wiley, New York, 2007, 367–422.
62. C. Curutchet, A. Bidon-Chanal, I. Soteras, M. Orozco and F. J. Luque, *J. Phys. Chem. B*. 2005, **109**, 3565–3574.
63. C. Curutchet, M. Orozco and F. J. Luque, *J. Comput. Chem.* 2001*,* **22***,* 1180–1193.
64. F. J. Luque, Y. Zhang, C. Alemán, M. Bachs, J. Gao and M. Orozco, *J. Phys. Chem*. 1996, **100**, 4269–4276.
65. F. J. Luque, M. Bachs, C. Alemán and M. Orozco, *J. Comput. Chem*. 1996, **17**, 806–820.
66. A. V. Marenich, C. P. Kelly, J. D. Thompson, G. D. Hawkins, C. C. Chambers, D. J. Giesen, P. Winget, C. J. Cramer and D. G. Truhlar, 2020, Minnesota Solvation Database (MNSOL) version 2012.
67. W. J. Dunn, M. G. Koehler and S. Grigoras, *J. Med. Chem.*, 1987, **30**, 1121–1126.
68. M.G. Koehler, S. Grigoras and W. J. Dunn, *Quant. Struct. Relationships,* 1988*,* **7***,* 150–159.
69. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H.P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J.C.

Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, Gaussian 16, (Revision C.01), Gaussian Inc., Pittsburgh, PA, 2016.

70. K. Lopez, S. Pinheiro and W. J. Zamora, *J. Comput. Aided Mol. Des.*, 2021, **35**, 923–931.
71. J. Plante and S. Werner, J. Cheminform., 2018, 10, 1–10.
72. H. F. Chen, Chem. Biol. Drug. Des., 2009, 74, 142–147.
73. A. Bahmani, S. Saaidpour and A. Rostami, Sci. Rep., 2017, 7, 1–14.
74. P. Patel, D. M. Kuntz, M. R. Jones, B. R. Brooks and A. K. Wilson, J. Comput. Aided Mol. Des., 2020, 34, 495–510.
75. V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. P. Feuston, J. Chem. Inf. Comput. Sci., 2003, 43, 1947–1958.
76. D. H. Kenney, R. C. Paffenroth, M. T. Timko and A. R. Teixeira, *J. Cheminform*., 2023, **15**, 1–14.
77. E. B. Lenselink and P. F. W. Stouten, *J. Comput. Aided Mol. Des*., 2021, **35**, 901–909.
78. M. Vallaro G. Ermondi, J. Saame, I. leito and G. Caron, B*ioorg. Med. Chem*., 2023, **81**, 117203.
79. C. J. Cramer and D. G. truhlar, Acc. Chem. Res. 2008, **41**, 760–768.
80. A. Klamt, B. Mennucci, J. Tomasi, V. Barone, C. Curutchet, M. Orozco and F. J. Luque, *Acc. Chem. Res*., 2009, **42**, 489–492.
81. W. J. Zamora, C. Curutchet, J. M. Campanera and F. J. Luque, *J. Phys. Chem. B*., 2017, **121**, 9868–9880.
82. W. J. Zamora, S. Pinheiro, K. German, C. Ràfols, C. Curutchet and F. J. Luque, *J. Comput. Aided Mol. Des*., 2020, **34**, 443–451.

83. A. Viayna, S. Pinheiro, C. Curutchet, F. J. Luque and W. J. Zamora, *J. Comput. Aided Mol. Des*., 2021, **35**, 803–811.
84. A. K. Chattah, R. Zhang, K. H. Mroue, L. Y. Pfund, M. R. Longhi, A. Ramamoorthy and G. Garnero, *Mol. Pharm*., 2015, **12**, 731–741.
85. J. Korać, N. Todorović, J. Zakrzewska, M, Zizic and I. Spasojevic, *Struct. Chem*., 2018, **29**, 1533–1541.
86. D. R. Silva, J. M. Silla, L. A. Santos, E. F. F. da Cunha and M. P. Freitas, *Mol. Inform*., 2019, **38**, 1800167.
87. G. Caron, M. Vallaro and M. Ermondi, *Drug Discov. Today Technol*. 2018, **27**, 65–70.
88. C. C. Bannan, G. Calabro, D. Y. Kyu and D. L. Mobley, *J. Chem. Theory Comput*., 2016, **12**, 4015–4024.
89. K. C. Chung and H. Park, *J. Comput. Aided Mol. Des.*, 2016, **30**, 1019–1033.
90. D. Santos-Martins, P. A. Fernandes and M. J. Ramos, *J. Comput. Aided Mol. Des*, 2016, **30**, 1079–1086.

807
808 *TOC Graphics*
809
810
811



812