Supporting Information

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[1,2,3,*], Antonio Viayna[4,5,6], Silvana Pinheiro[1,2], Carles Curutchet[6,7], Laia Bisbal[5,8], Rebeca Ruiz[9], Clara Ràfols[5,8,*], F. Javier Luque[4,5,6,*]

[1] CBio[3] Laboratory, School of Chemistry, University of Costa Rica, San Pedro, San José, Costa Rica.

[2] Laboratory of Computational Toxicology and Artificial Intelligence (LaToxCIA), Biological Testing Laboratory (LEBi), University of Costa Rica, San Pedro, San José, Costa Rica.

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

[4] 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.

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

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

[7] 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.

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

[9] 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)

Table S1 Recommended volumes of water and toluene according to the estimated log$P_{tol/w}$ value for multisets using Pion SiriusT3 (Pion Inc.).

Estimate log$P_{tol/w}$ value	Number of Assays	Toluene volume (mL)			Water volume (mL)
		1st assay	2nd assay	3rd assay	
Below -1	1	1.5			0.75
Below 0	2	0.5	1.0		0.75
Between 0 and 1	3	0.1	0.4	0.75	0.75
Between 1 and 2	3	0.05	0.2	1.0	0.75
Between 2 and 3	3	0.02	0.1	1.0	0.75
Above 3	3	0.01	0.1	1.0	0.75

Table S2 Experimental log$P_{tol/w}$ values for a subset of compounds taken from the literature [28] and determined in this work.

Compound	Reported	This work	Compound	Reported	This work
2-iodophenol	1.5	1.6	2-naphthoic acid	1.8	1.9
3-chlorophenol	1.0	1.0	3,5-dimethoxyphenol	0.7	0.7
4-hydroxybenzoic acid	-1.4	-0.8	4-nitrophenol	-0.2	-0.1
4-phenylbutylamine	1.7	1.7	aniline	0.9	0.8
chlorpromazine	6.6	6.6	desipramine	4.0	4.2
diclofenac	3.0	3.1	diltiazem	3.5	3.4
diphenhydramine	3.8	3.8	eserine	1.1	1.0
flumequine	1.5	1.5	fluoxetine	4.9	4.9
flurbiprofen	3.0	3.1	haloperidol	3.1	3.2
ibuprofen	2.8	2.9	indomethacin	3.0	3.5
lidocaine	2.4	2.3	metoprolol	1.1	1.1
naproxen	2.3	2.3	nifuroxime	0.0	0.0
papaverine	3.1	3.2	penbutolol	4.1	4.1
phenol	0.1	0.1	phenylacetic acid	0.1	-0.3
procaine	1.6	1.3	propanolol	2.5	2.6
tetracaine	3.4	3.3	thymol	2.1	2.2
tramadol	2.7	2.7	verapamil	4.8	5.0

Table S3 Statistical parameters of the comparison between experimental and predicted $logP_{tol/w}$ values for the training, test, and external set using the regression models with the initial 79 descriptors.

Method	r^2			RMSE		
	Training	Test	External	Training	Test	External
MLR	0.97	0.86	0.82	0.34	0.64	1.31
RFR	0.97	0.82	0.89	0.36	0.78	0.86

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

Table S4 Statistical results of the *k*-fold cross-validation (k = 5) performed for the MLR and RFR models.[a]

Model	RMSE	r^2	MUE
MLR	1.04	0.70	0.79
RFR[b]	1.04	0.73	0.79

[a] RMSE: root-mean square error (in log *P* units); r^2: squared coefficient of determination; MUE: mean unsigned error in cross-validation analysis.

[b] The number of variables randomly sampled at each split was 3 and 187 trees were used.

Table S5 Selected experimental physicochemical properties (at 298 K) of toluene and benzene.

Property	Benzene	Toluene
Molar mass (g mol^{-1})	78.11	92.14
Density (g mL^{-1})[a]	0.87	0.86
Dielectric constant[a]	2.27	2.38
Dipole (D)	0.00	0.33
Quadrupolar components (D Å)[b]	(4.0, 4.0, -8.0)	(-7.4, 3.9, 3.5)
Polarizability ($Å^3$)	10.0	11.9
Ionization energy (eV)	9.2	8.8

[a] Used in the parametrization of the IEF-PCM/MST continuum solvation model.

[b] Determined at the MP2/aug-cc-pVTZ level.

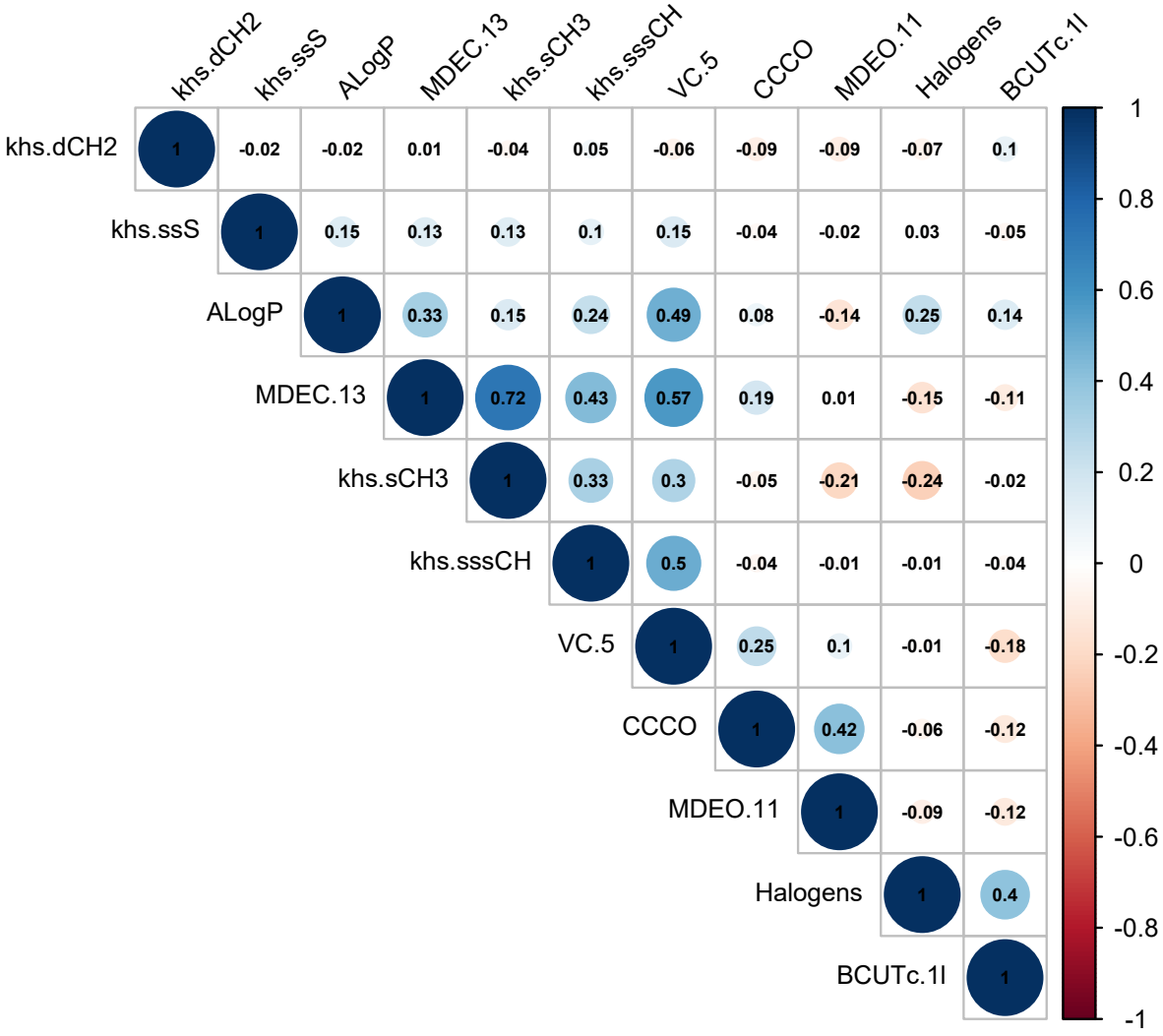

Fig. S1 Matrix of Pearson's correlation coefficient (*r*) of the 11 predictors used in the derivation of the ML models for the subset of compounds included in the training set.

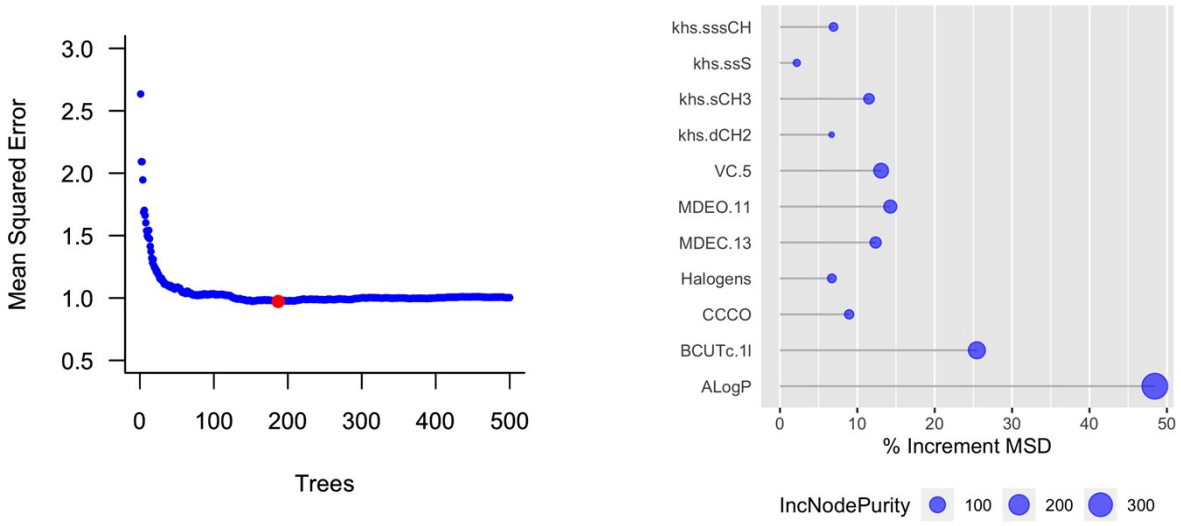

Fig. S2 (Left) Representation of the number of trees as a function of the mean square error. The minimum mean squared error was found for a value of 187 trees (red point). (Right) Variable importance in the RFR model for the training set (n = 214). The importance was evaluated through the mean decrease accuracy expressed as percentage of increment in the mean square deviation (MSD) if the variable is dropped, and using the Gini impurity index (IncNodePurity), which represents the total decrease in residual sum of squares when splitting on a variable averaged over all trees.

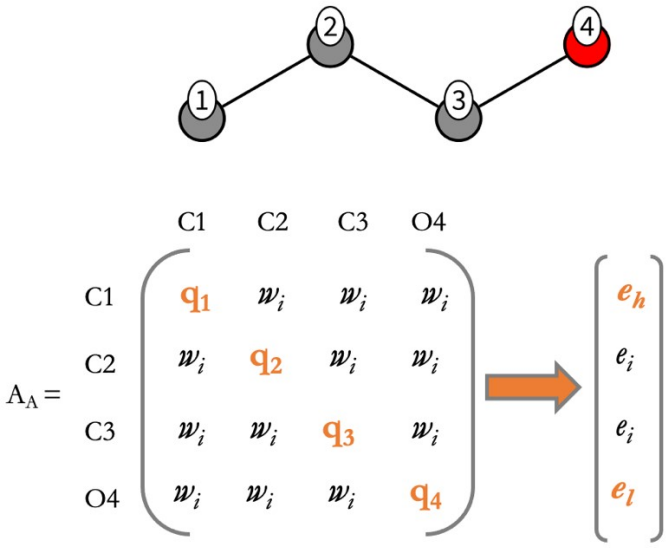

Fig. S3 BCUT class descriptor - based on Burden's (B) original suggestion and CAS's (C) validation followed of extensions proposed by Pearlman at the University of Texas (UT), which resulted in what is now referred to as the BCUT approach. This descriptor works on an adjacency matrix A_A of the hydrogen-depleted molecule (e.g., propan-1-ol) where for the specific descriptor BCUTc.1l, the Gasteiger-Marsili empirical atomic partial charges (q_i) are placed on the diagonal of the matrix and the off-diagonal elements (w_i) are weighted according to the nominal bond-type (0.1 times the bond-type if the two atoms are bonded, 0.01 to terminal bonds to the last atom in a chain, and 0.001 if the two atoms are not bonded). Eigenvalues (e_i) are obtained by solving the matrix, thus representing a one-dimensional chemistry-space; however, the lowest (e_l) and highest (e_h) eigenvalues are the only ones returned as descriptors because they reflect the most differing information.

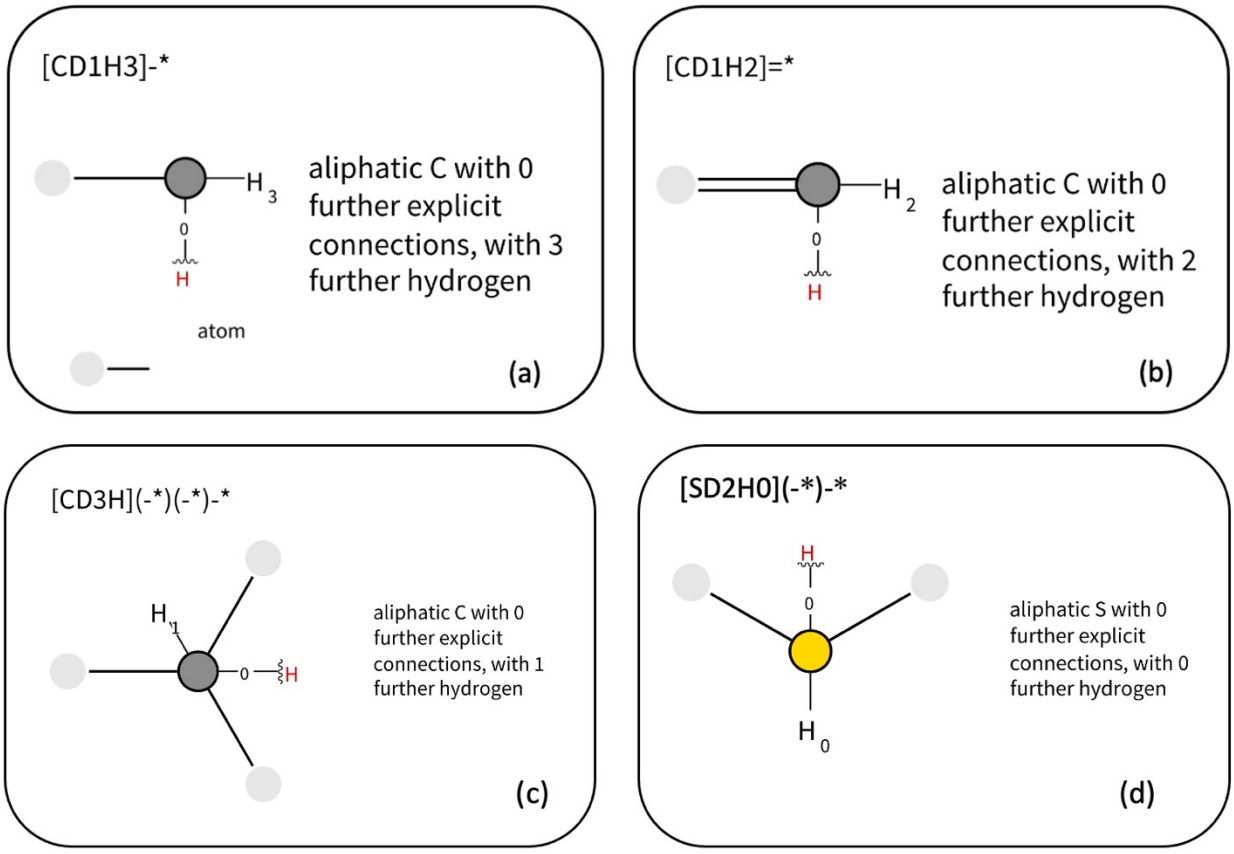

Fig. S4 Kier-Hall smarts descriptors (a) khs.sCH3, (b) khs.dCH2, (c) khs.sssCH, and (d) khs.ssS, used as predictors in the study for the ML models.

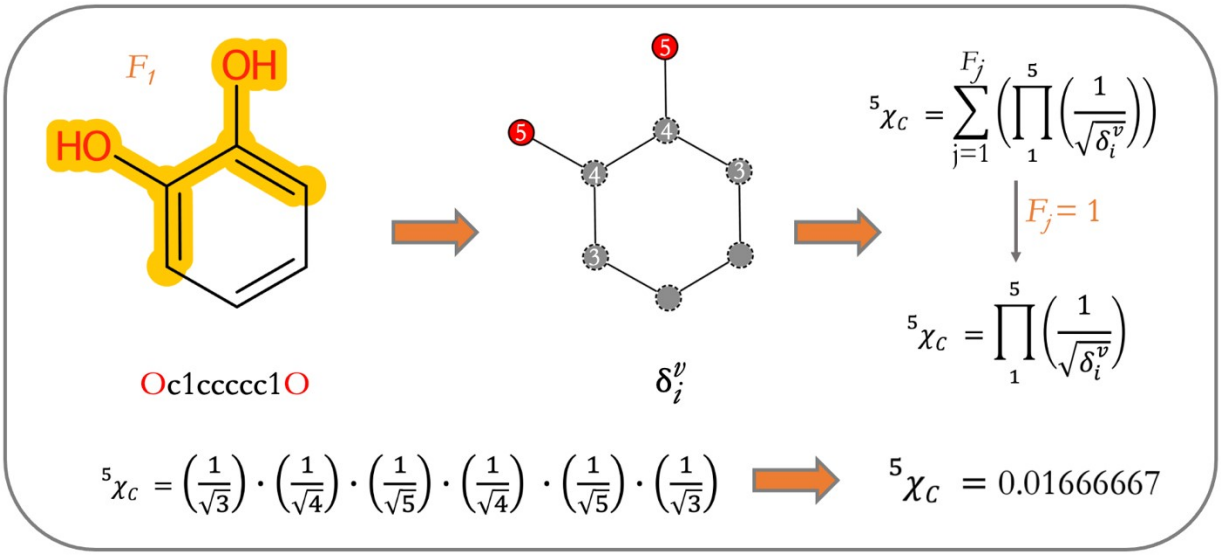

Fig. S5 The valence Kier-Hall Chi cluster descriptor of order 5 (VC.5) exemplified for catechol. The fragment F1 (orange) represents the only cluster of order 5 in the molecule. The delta valence value ($^{\delta^v_i}$) for each atom (i) is calculated as $^{\delta^v_i} = Z^v_i - h_i$, where Z^v_i is the number of valence electrons for the atom (i) and h_i stands for the number of hydrogen atoms bonded to the atom (i). Finally, the valence Kier-Hall Chi cluster descriptor of order 5 ($^5\chi_c$) can be determined using the δ^v_i values and the formula shown in the figure.

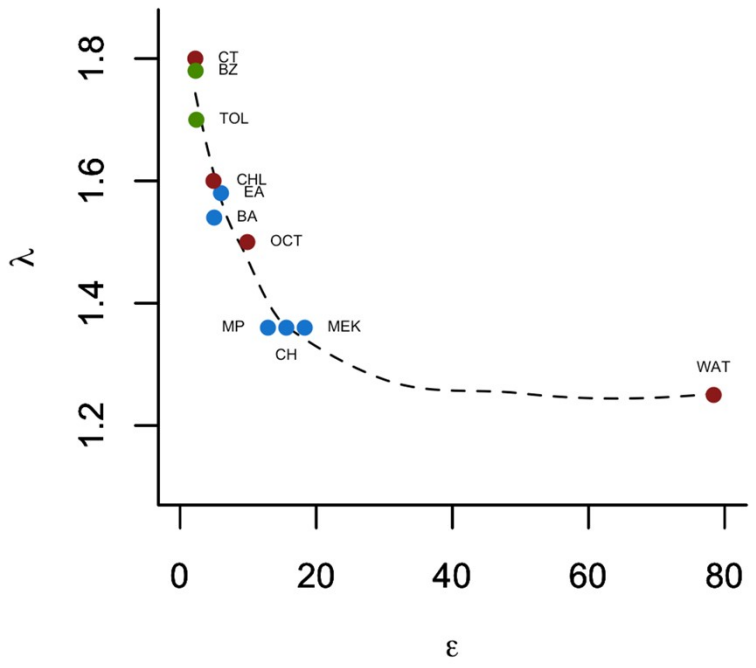

Fig. S6 Representation of the solvent-dependent scaling factor (λ) used to scale the atomic radii to build up the cavity of the solute as a function of the dielectric constant (ε) of the solvent. Values optimized for water (WAT), n-octanol (OCT), chloroform (CHL) and carbon tetrachloride (CT) in the IEF-PCM/MST model are shown as red circles, the interpolated λ values used in the SAMPL8 log*P* challenge (manuscript under preparation) for ethyl acetate (EA), butyl acetate (BA), methyl ethyl ketone (MEK), cyclohexanone (CH) and 4-methyl-2-pentanone (MP) as blue circles, and the interpolated λ values for toluene (TOL) and benzene (BZ) as green circles.