Supporting Information for Integrating Multiscale and Machine Learning Approaches toward the SAMPL9 LogP Challenge

Michael R. Draper, Asa Waterman, Jonathan E. Dannatt*, and Prajay Patel*
Corresponding Email: jdannatt@udallas.edu, pmpatel@udallas.edu

Table of Contents

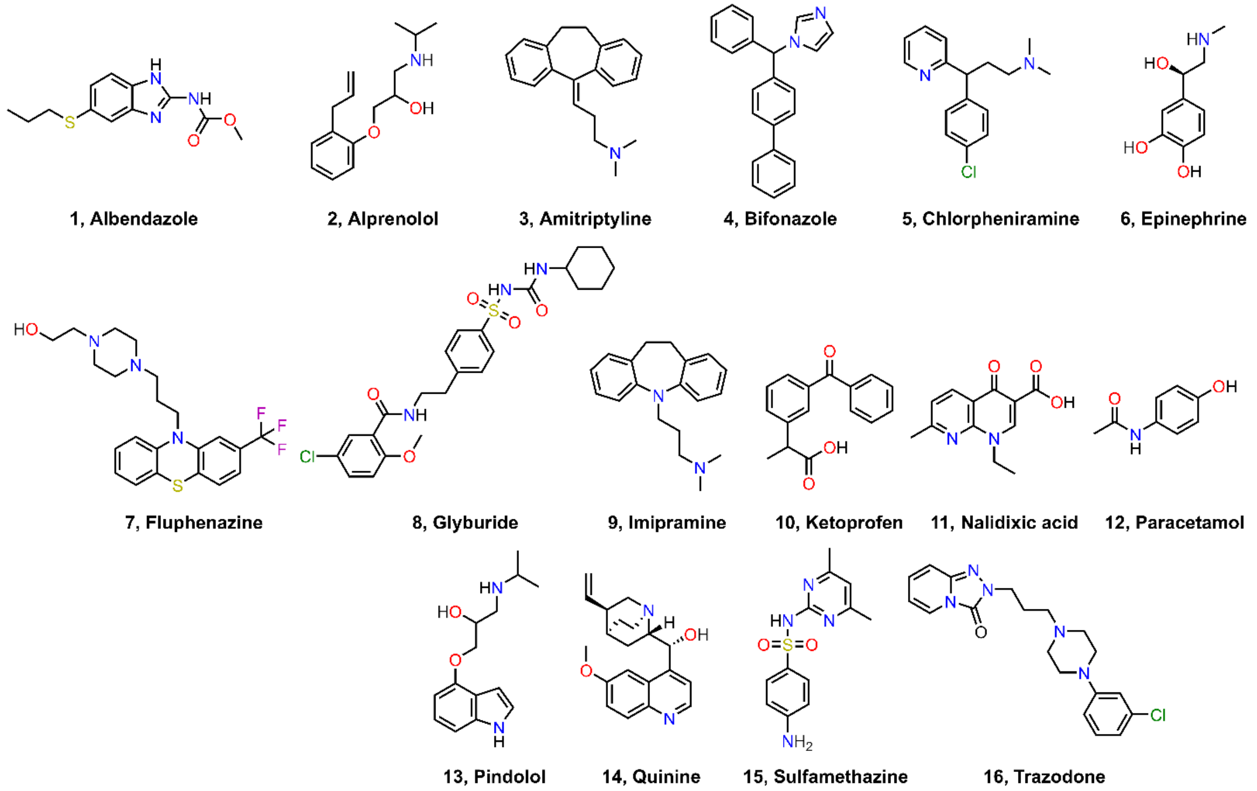

Figure S1. 2D structures of the sixteen challenge molecules. Same as Figure 1.

Table S1. The given labeling scheme and SMILES strings of the sixteen challenge molecules as part of the SAMPL9 LogP challenge.

Label	Molecule	SMILES
1	Albendazole	CCCSc1ccc2c(c1)[nH]c(n2)NC(=O)OC
2	Alprenolol	CC(C)NCC(O)COc1ccccc1CC=C
3	Amitriptyline	CN(C)CCC=C2c1ccccc1CCc3ccccc23
4	Bifonazole	c1ccc(cc1)C(c2ccc(cc2)c3ccccc3)n4ccnc4
5	Chlorpheniramine maleate salt	CN(C)CCC(c1ccc(Cl)cc1)c2ccccn2
6	Epinephrine	CNC[C@H](O)c1ccc(O)c(O)c1
7	Fluphenazine dihydrochloride	OCCN4CCN(CCCN2c1ccccc1Sc3ccc(cc23)C(F)(F)F)CC4
8	Glyburide	COc1ccc(Cl)cc1C(=O)NCCc2ccc(cc2)S(=O)(=O)NC(=O)NC3CCCCC3
9	Imipramine hydrochloride	CN(C)CCCN2c1ccccc1CCc3ccccc23
10	Ketoprofen	CC(C(O)=O)c1cccc(c1)C(=O)c2ccccc2
11	Nalidixic acid	CCn1cc(C(O)=O)c(=O)c2ccc(C)nc12
12	Paracetamol	CC(=O)Nc1ccc(O)cc1
13	Pindolol	CC(C)NCC(O)COc1cccc2[nH]ccc12
14	Quinine	COc4ccc3nccc(C(O)C1CC2CCN1CC2C=C)c3c4
15	Sulfamethazine	Cc2cc(C)nc(NS(=O)(=O)c1ccc(N)cc1)n2
16	Trazodone hydrochloride	Clc1cccc(c1)N4CCN(CCCn3nc2ccccn2c3=O)CC4

Online Resources

Software	Description	Website
ORCA	Using a vertical solvation approach with DFT to compute logP$_{o/w.}$	https://www.orcasoftware.de/tutorials_orca/prop/CPCM.html
GROMACS	Tutorial for computing the free energy of solvation of ethanol in water.	https://tutorials.gromacs.org/free-energy-of-solvation.html https://tutorials.gromacs.org/awh-free-energy-of-solvation.html

Quantum Mechanical (QM) Approaches

Density Functional Theory (DFT) Calibration

Fifteen functionals from five developer families were included in the calibration of density functionals to examine the effects of density functional tier and parameterization on predicting logP$_{tol/w}$. These are shown in Table S1. The correlation consistent basis sets (cc-pVDZ, cc-pVTZ, cc-pVQZ)[1] were used to compare the effect of basis sets on logP$_{tol/w}$. The recommended cc-pV(n+d)Z basis sets were used for all S and Cl atoms.[2] Various extrapolation schemes have been developed to extrapolate electronic energies to the complete basis set (CBS) limit using Dunning's correlation consistent basis sets, including three-point extrapolation schemes based on the ζ-level of the basis set (Peterson)[3] and two-point extrapolation incorporating the maximum angular momentum of the basis set (Schwartz).[4–7] For this work, a mixed Peterson-Schwartz extrapolation scheme (PS3) that averages the Peterson (P) three-point with the Schwartz-3 (S3) two-point extrapolation with triple- and quadruple-zeta level basis sets or PS3(TQ) is used for all extrapolations to the CBS limit.[8,9] This mixed extrapolation scheme has been shown to correct the over- and underestimation of the CBS limit of the S3 and P schemes, respectively, due to their respective rates of convergence.

Table S2. Density functionals used in this study.

Functional Family	Functional	Tier[a]	% Exact Exchange
BXLYP	BLYP[10,11]	GGA	0%
	B3LYP[12]	H-GGA	20%
	BHandHLYP[13]	H-GGA	50%
PBE	PBE[14,15]	GGA	0%
	PBE0 [14–16]	H-GGA	25%
	revPBE0[14,15]	H-GGA	25%
TPSS	TPSS[17]	M-GGA	0%
	TPSSh[17,18]	HM-GGA	10%
	TPSS0[19]	HM-GGA	25%
Minnesota	M06L[20]	M-GGA	0%
	M06 [21]	HM-GGA	27%
	M06-2X[21]	HM-GGA	54%
ω	ωB97 [22–25]	Range-separated hybrid	0%
	ωB97X [22–25]	Range-separated hybrid	15.7%
	ωB97X-V [22–25]	Range-separated hybrid	16.7%

[a] GGA = generalized gradient approximation, H-GGA (hybrid-GGA), M-GGA (meta-GGA), HM-GGA (hybrid meta GGA).

Table S3. The mean unsigned error (MUE) and standard deviation (σ) of each basis set and functional combination between the calculated and experimental logP$_{tol/w}$.

Functionals	cc-pVDZ	cc-pVTZ	cc-pVQZ	cc-pV∞Z
BLYP	1.41±0.86	1.15±1.03	1.13±1.01	1.13±0.99
B3LYP	1.09±0.92	1.20±0.96	1.23±0.93	1.24±0.92
BHandHLYP	1.14±0.97	1.47±1.02	1.49±1.02	1.50±1.02
M06L	1.49±1.09	1.21±1.07	1.20±1.07	1.19±1.08
M06	1.32±1.07	1.18±0.97	1.18±0.96	1.17±0.95
M06-2X	1.12±0.98	1.27±0.93	1.25±0.91	1.24±0.89
PBE	1.33±0.82	1.00±0.97	1.00±0.95	1.00±0.93
PBE0	0.96±0.90	1.13±0.88	1.14±0.87	1.15±0.86
REVPBE0	0.99±0.78	1.12±0.90	1.13±0.88	1.14±0.87
TPSS	1.31±0.83	1.08±1.00	1.06±0.98	1.06±0.96
TPSSh	1.15±0.86	1.10±0.96	1.10±0.94	1.09±0.93
TPSS0	0.97±0.93	1.19±0.92	1.20±0.90	1.20±0.89
ωB97	1.13±0.94	1.17±0.88	1.19±0.88	1.21±0.87
ωB97X	1.09±0.93	1.20±0.89	1.21±0.89	1.23±0.88
ωB97X-V	1.13±0.92	1.16±0.88	1.18±0.87	1.18±0.86

Table S4. The mean unsigned error (MUE) and standard deviation (σ) when varying a functional and keeping a basis set constant for the deviations between the calculated and experimental logP$_{tol/w}$ values.

Basis Set	MUE±σ
cc-pVDZ	1.18±0.91
cc-pVQZ	1.18±0.92
cc-pVTZ	1.18±0.93
cc-pV∞Z	1.18±0.91

One of the main goals of the DFT calibration was to find an optimal method/basis set combination that would generally apply to computing logP$_{tol/w}$, based on the mean unsigned error (MUE), and standard deviation (σ). Based on Figure S2, cc-pV∞Z had the lowest change in logP due to removing basis set incompleteness error. However, in Figure S2, and in Table S4, the MUE when comparing all basis sets was the same. Therefore, regardless of functional choice, changing the basis set would have minimal effect on the logP value. cc-pVTZ was chosen for the submission trial because it required less processing power and time than cc-pVQZ and cc-pV∞Z. Triple-ζ quality basis sets provide a compromise between computational cost and accuracy, and so, even though using cc-pVDZ would save CPU time when considering numerous DFT calculations, cc-pVDZ yielded the highest MUE for the chosen functional (PBE). Table S3 showed that the functional PBE had the lowest MUE and was chosen to be the function for the submission trial. In Table S3, the hybrid functionals yielded a higher MUE and σ than the GGA functionals.

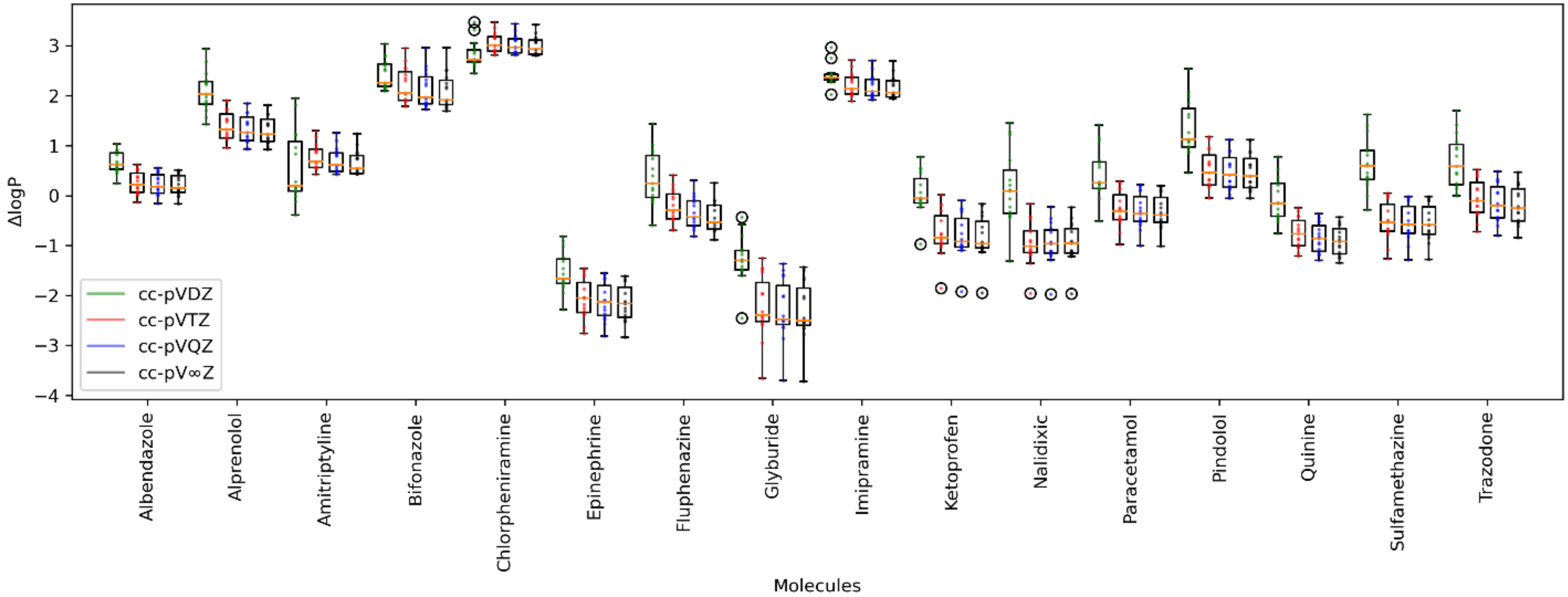

Figure S2. A box and whiskers plot that compares the signed errors (ΔlogP) of the calculated logP$_{tol/w}$ from the experimental logP$_{tol/w}$ values for the sixteen challenge molecules for each basis set. The box highlights the distribution of logP when selecting different density functionals. The orange line represents the average value in each box, and the black circles represent outliers. The basis sets examined are cc-pVDZ (green), cc-pVTZ (red), cc-pVQZ (blue), and cc-pV∞Z (black).

DLPNO-Solv-ccCA

The correlation consistent Composite Approach (ccCA) was developed by Wilson and coworkers as an alternative to the Gaussian-n (G*n*) composite methods. Several variants have emerged over the years to describe various chemical phenomena across the periodic table including transition metal chemistry, solvated species, and lanthanides.[26] The DLPNO-Solv-ccCA[27] method (Table S5) is a combination of the DLPNO-ccCA[28] and Solv-ccCA[29] methods targeting main group thermochemistry, and was developed as part of the SAMPL6 competition to predict the logP$_{o/w}$.[27] Composite approaches use stepwise additive corrections to approximate a higher level of theory as shown in Equation S1.

$$E_{DLPNO-Solv-ccCA} = E_{ref} + \Delta CC + \Delta CV + \Delta SR + ZPE \ \#(S1)$$

Table S5. A Schematic for DLPNO-Solv-ccCA.[27]

Geometry Optimization	B3LYP-D3/cc-pVTZ (gas)
RIJCOSX-HF/CBS	RIJCOSX-HF/aug-cc-pVDZ SMD(solvent) RIJCOSX-HF/aug-cc-pVTZ SMD(solvent) RIJCOSX-HF/aug-cc-pVQZ SMD(solvent) $E(n) = E_{CBS} + Be^{-1.63n}$
DLPNO-MP2/CBS	DLPNO-MP2/aug-cc-pVDZ SMD (solvent) DLPNO-MP2/aug-cc-pVTZ SMD (solvent) DLPNO-MP2/aug-cc-pVQZ SMD (solvent) $E_P(x) = E_{CBS} + Be^{-(x-1)} + Ce^{-(x-1)^2}$ $E_{S3}(l_{max}) = E_{CBS} + \frac{B}{\left(l_{max} + \frac{1}{2}\right)^4}$
ΔCC	DLPNO-CCSD(T)/cc-pVTZ SMD (solvent) – DLPNO-MP2/cc-pVTZ SMD (solvent)
ΔCV	DLPNO-MP2/aug-cc-pCVTZ SMD (solvent) – DLPNO-MP2/aug-cc-pVTZ SMD (solvent)
ΔSR	DLPNO-MP2/cc-pVTZ-DK SMD (solvent) – DLPNO-MP2/cc-pVTZ SMD (solvent)
ZPE	B3LYP-D3/cc-pVTZ frequencies scaled by 0.9890

Solvent is either water or toluene. The reference energy is obtained from the Hartree-Fock (HF) energy extrapolated to the complete basis set (CBS) limit and an extrapolation of the DLPNO-MP2 correlation energy using the PS3(TQ) extrapolation scheme. The correlation correction (ΔCC), accounting for higher level electron correlation, core-core/core-valence (ΔCV) correlation energies, and scalar relativistic (ΔSR) corrections are added to the reference energy to yield the total energy. The correlation consistent auxiliary basis sets were used for the DLPNO methods and the def2 auxiliary basis set was used for the Hartree-Fock coulomb and exchange integral computations.

Molecular Mechanics

Table S6. The decoupled 25-point λ path used to compute the solvation free energy in each solvent. 0.00 means that the molecule is fully decoupled and does not interact with its surroundings, and 1.00 means that the interactions are turned on as noted in Ref 30.

vdW Interactions	Coulomb Interactions
0.00	0.00
0.10	0.00
0.20	0.00
0.22	0.00
0.24	0.00
0.26	0.00
0.28	0.00
0.30	0.00
0.33	0.00
0.36	0.00
0.40	0.00
0.45	0.00
0.50	0.00
0.60	0.00
0.70	0.00
0.80	0.00
0.90	0.00
1.00	0.00
1.00	0.17
1.00	0.34
1.00	0.48
1.00	0.62
1.00	0.75
1.00	0.88
1.00	1.00

Data Science and Unsupervised Machine Learning

Unsupervised machine learning was used to cluster conformer structures using pairwise nucleus-nucleus distance matrices as a high dimensional input that was reduced via principal component analysis. The variation caused by creating pairwise matrices should provide more information into potential clustering of data by structural conformations within a given solvent. 5000 structures extracted from the MD simulations of each molecule in toluene and water were split into randomized training sets containing 2000, 3000, 4000, and 5000 structures to perform a five-fold cross-validation of the number and composition of the clusters predicted by the unsupervised machine learning algorithms (K-means, Gaussian Mixture Models). Different training set sizes were utilized to balance the density of each cluster while retaining the information of structural conformations as including more data points into the clusters may provide information into cluster density that may be attributed to transient species but may be considered intermediates. Figure S3 demonstrates this concept for **5**, in which sampling 2K points and 4K points effectively identifies similar clusters across the first two principal components. The 4K sampling adds more datapoints to the lower right region near (0.6, -0.2), which would translate the cluster mean relative to its location in the 2K sampling size, and therefore, which structures were sampled as part of the ML method presented in Table 3.

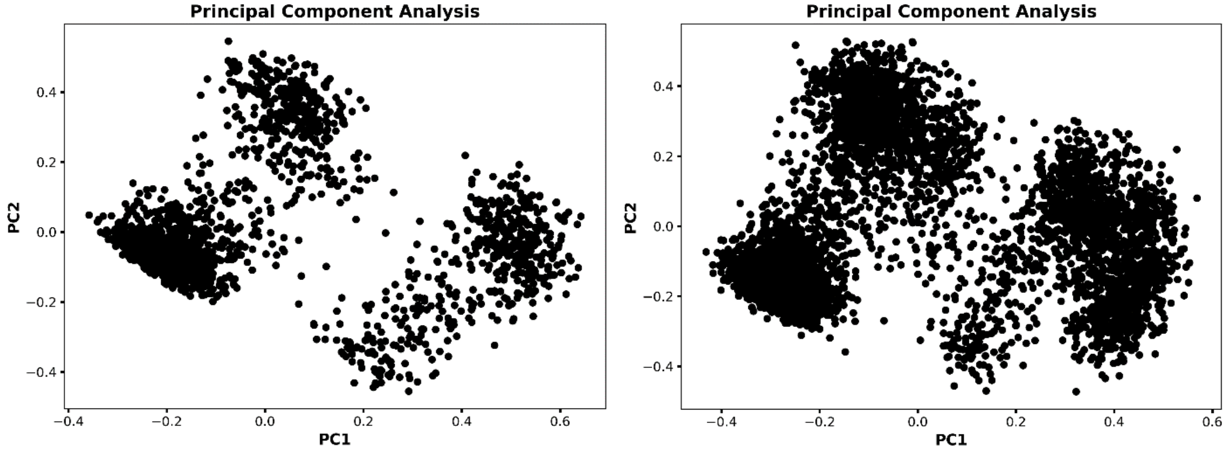

Figure S3. The first two principal components of the distance matrix descriptor used for the ML method when using a randomized 2K sample size (left) and 4K sample size (right) for **5**.

The averaged silhouette scores after the five-fold cross validation for each training set size are reported in Table S7. The silhouette score that was closest to the average silhouette score across all training sets was the criteria for selecting the number of structures to use in the final clustering, shown in Table 3. Since the sample size of 4K structures had the most silhouette scores closest to the average across all molecules, the clustering results from the Gaussian Mixture Models when using 4K structures were utilized in the final combined results.

Based on the logP values from the 2K analysis, the cluster assignment might be wrong. So, more sample sizes were created. Table S7 shows that a 4K sample size has the closest silhouette scores to the average score for each molecule. Silhouette scores are important because they find the optimum cluster assignment for each molecule. Table S8 shows how the 4K $\log P_{tol/w}$ values differ from the experimental $\log P_{tol/w}$ values. In Table S8, using the 4K sample size was marginally better than the 2K sample size (0.01 $\log P_{tol/w}$ units). An explanation for this is the incorrect cluster assignment. Since cluster assignment changes as a sample size changes, then the cluster assignment impacts the $\log P_{tol/w}$ value more than the sample size itself. Because of the complex nature of chemical systems, they are difficult to model and there are often no systematic ways to determine the accuracies of blind predictions. Despite the change in sample size, the difference between $\log P_{tol/w}$ values were insignificant. This would suggest that there is some bias in the assumptions of sampling based on sample size and the structures used to represent the clusters. Given the statistical nature of chemistry, many of the structures generated through these dynamical simulations may not be true equilibrium structures or observable through standard experimental characterization techniques like NMR or GC/MS. Therefore, human intuition is still needed as part of the research process when building the machine learning methods that attempt to serve as “black-box” methods.

Table S7. The averaged silhouette scores for each molecule in four different sample sizes following the five-fold cross validation. The average of the averaged silhouette scores and the number of clusters predicted using a 4K sample size is shown.

Label	5K	4K	3K	2K	Average Silhouette Score	Clusters
1	0.346	0.349	0.351	0.350	0.349	2
2	0.466	0.462	0.463	0.462	0.463	4
3	0.578	0.575	0.570	0.569	0.573	2
4	0.445	0.451	0.455	0.457	0.452	5
5	0.594	0.592	0.589	0.589	0.591	5
6	0.502	0.503	0.500	0.506	0.503	4
7	0.426	0.427	0.428	0.427	0.427	4
8	0.476	0.477	0.475	0.477	0.476	3
9	0.641	0.642	0.640	0.641	0.641	8
10	0.611	0.611	0.622	0.604	0.612	8
11	0.636	0.630	0.622	0.613	0.625	2
12	0.441	0.442	0.443	0.450	0.444	6
13	0.495	0.493	0.491	0.491	0.493	5
14	0.290	0.293	0.298	0.323	0.301	4
15	0.579	0.575	0.573	0.530	0.564	2
16	0.556	0.555	0.553	0.552	0.554	8

Table S8. The calculated logP$_{tol/w}$ values and the mean unsigned error (MUE) for the ML method using a 2K and 4K sample size. The best results of the two sample sizes are combined to mitigate the effect of sampling bias and are also shown in Table 3 as the ML method.

Label	2K sample size	4K sample size	Exp
1	-1.73	1.79	3.76
2	-0.54	1.59	2.40
3	0.83	-1.71	5.51
4	-1.33	-1.74	5.47
5	-0.73	4.33	3.61
6	0.94	-2.72	-1.23
7	2.34	0.31	4.37
8	-1.83	4.95	2.79
9	6.49	-2.33	5.05
10	0.48	-0.55	2.47
11	-0.41	2.65	1.46
12	-3.53	-1.18	-1.59
13	0.09	-1.40	0.36
14	-0.84	0.68	1.41
15	-3.14	-5.09	-0.74
16	3.50	2.91	3.77
MUE	2.84	2.83	

Spartan20 (S20) Method

The conformations generated via the Spartan '20 program uses the Merck Molecular Force Field (MMFF) to performs its conformation search. The Boltzmann-weighted PBE/cc-pVTZ energies of the conformations are scaled relative to the minimum where conformations with energies closer to the minimum energy are weighted more than conformations with higher energies. As shown in Figure S4, the shape of the violin plots indicates the notable changes in the distributions of the weights in the Boltzmann scheme versus the RMSD scheme. This observation is most likely due to the stronger effect of the solvent on the electronic energies than the RMSD of the structures with respect to the optimized molecule in each solvent. The overall larger change in the Boltzmann weights indicates that the larger range of the conformers generated were the cause of the larger errors associated with predicted logP values rather than the method of Boltzmann weighting. For the RMSD violin plots, there were no major changes to the distributions with respect to each solvent, indicating the optimized structures in each solvent were similar in the 3D geometry.

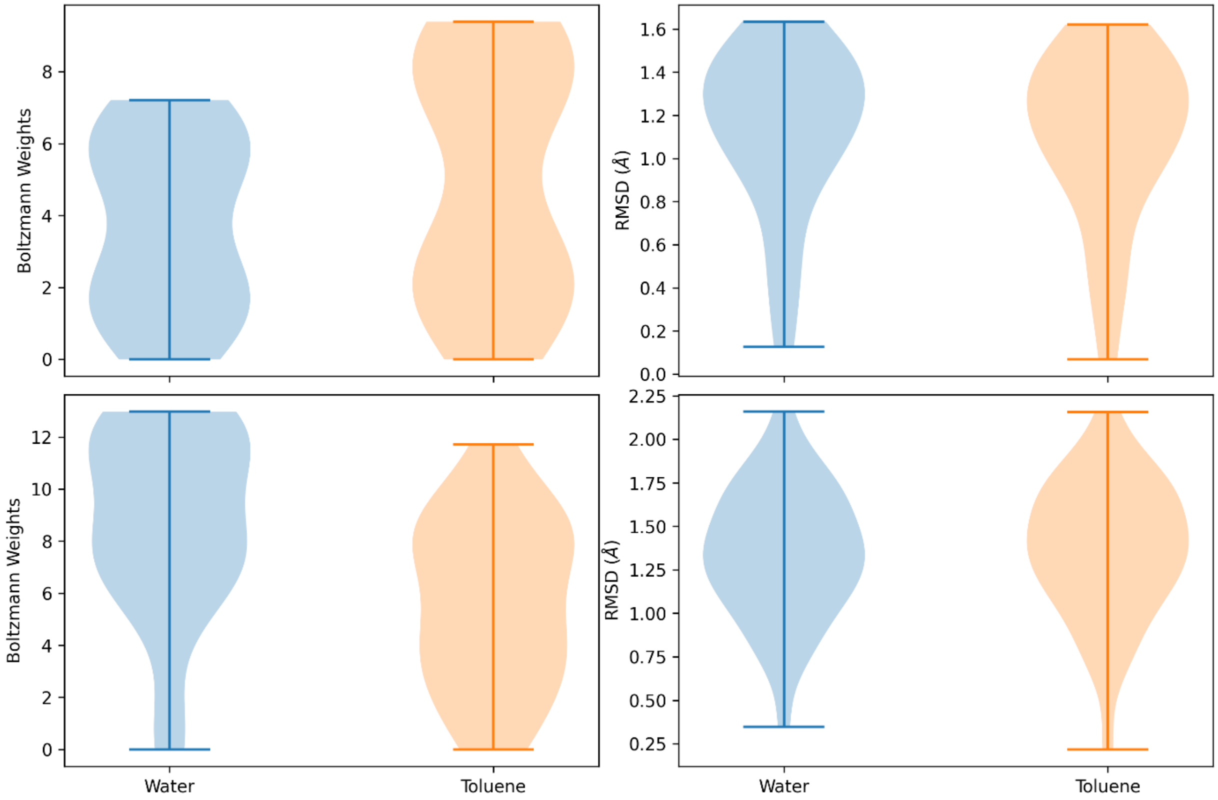

Figure S4. Violin plots showing the weighting distributions from the Boltzmann-weighted scheme (left) and the RMSD-weighted scheme (right) for **1** (top) and **5** (bottom).

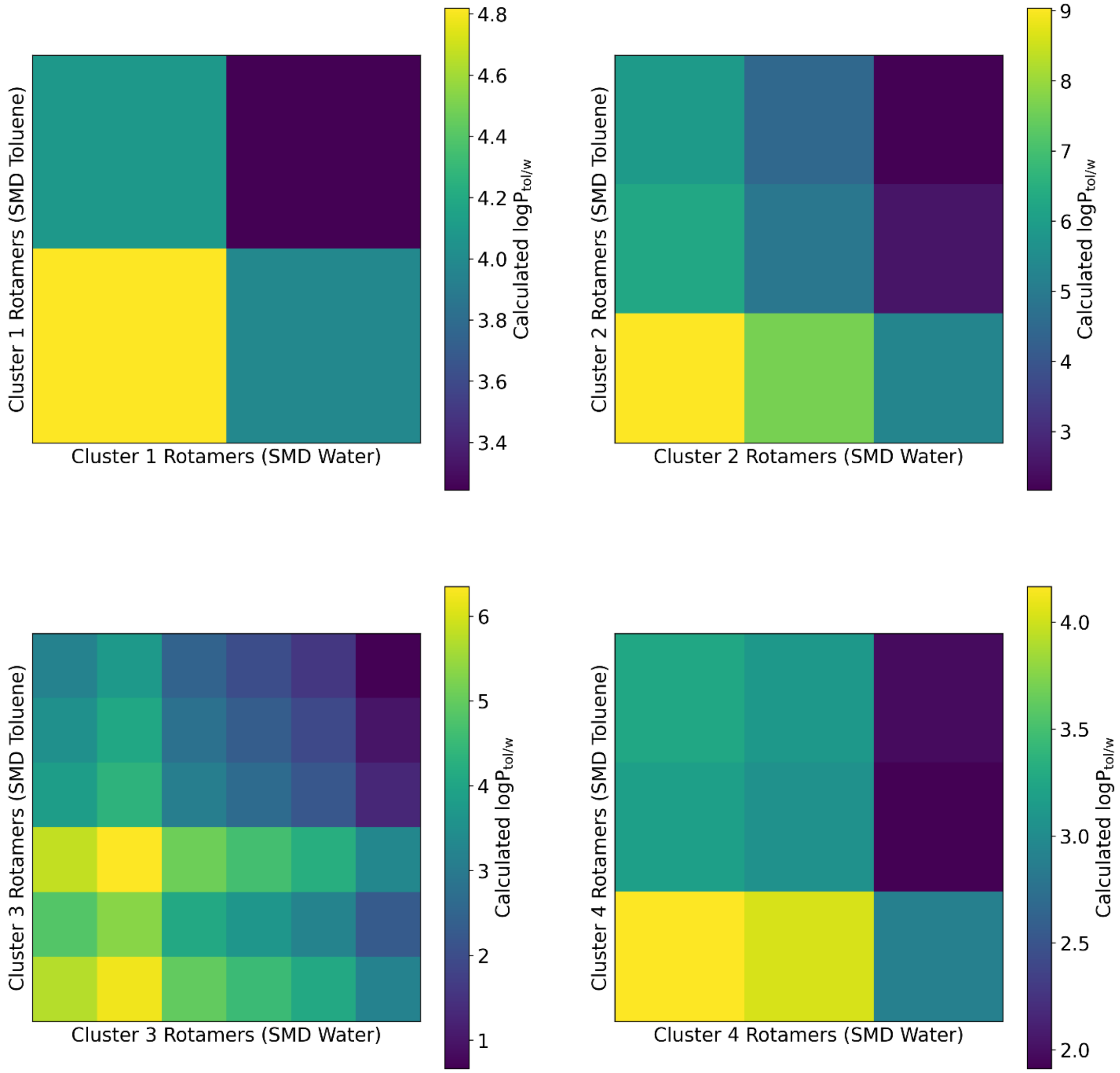

Figure S5. The logP$_{tol/w}$ distribution of the clusters predicted via the Gaussian Mixture Models for **5** using the S20 method to generate the conformers and calculate the logP$_{tol/w}$. Each square in the heat map represents a different combination of the conformers assigned to particular clusters to compute logP$_{tol/w}$.

As seen in Figure S5, the range of calculated logP$_{tol/w}$ values varies between the different clusters, showing no clear indication that the cluster assignments correlate to the structural differences that cause significant changes in the calculated logP$_{tol/w}$.

Table S9. The weighting of the peaks used in the Mixed submission based on the histogram of calculated $\log P_{tol/w}$ values.

Label	# peaks	% peak 1	% peak 2	% peak 3	% peak 4	% peak 5	Total
1	2	0.0381	0.9619	0	0	0	1
2	1	1	0	0	0	0	1
3	1	1	0	0	0	0	1
4	1	1	0	0	0	0	1
5	1	1	0	0	0	0	1
6	1	1	0	0	0	0	1
7	1	1	0	0	0	0	1
8	1	1	0	0	0	0	1
9	1	1	0	0	0	0	1
10	1	1	0	0	0	0	1
11	4	0.1315	0.6336	0.2299	0.0050	0	1
12	1	1	0	0	0	0	1
13	1	1	0	0	0	0	1
14	3	0.1279	0.7197	0.1524	0	0	1
15	5	0.0180	0.1551	0.5871	0.2260	0.0138	1
16	1	1	0	0	0	0	1

XYZ files

All XYZ coordinate files of the optimized molecules at the RIJCOSX-B3LYP-D3BJ/cc-pVT(+d)Z level of theory in both implicit solvents are provided in a separate zipped folder. All the molecules were assumed to be neutral charge with all electrons paired.

References

1 T. H. Dunning Jr, Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen, *J. Chem. Phys.*, 1989, **90**, 1007–1023.

2 T. H. Dunning Jr, K. A. Peterson and A. K. Wilson, Gaussian basis sets for use in correlated molecular calculations. X. The atoms aluminum through argon revisited, *J. Chem. Phys.*, 2001, **114**, 9244–9253.

3 K. A. Peterson, D. E. Woon and T. H. Dunning Jr, Benchmark calculations with correlated molecular wave functions. IV. The classical barrier height of the H+H$_2$→H$_2$+H reaction, *J. Chem. Phys.*, 1994, **100**, 7410–7415.

4 W. Kutzelnigg and J. D. Morgan, Rates of convergence of the partial-wave expansions of atomic correlation energies, *J. Chem. Phys.*, 1992, **96**, 4484–4508.

5 J. M. L. Martin, Ab initio total atomization energies of small molecules — towards the basis set limit, *Chem. Phys. Lett.*, 1996, **259**, 669–678.

6 T. Helgaker, W. Klopper, H. Koch and J. Noga, Basis-set convergence of correlated calculations on water, *J. Chem. Phys.*, 1997, **106**, 9639–9646.

7 A. Halkier, T. Helgaker, P. Jørgensen, W. Klopper, H. Koch, J. Olsen and A. K. Wilson, Basis-set convergence in correlated calculations on Ne, N$_2$, and H$_2$O, *Chem. Phys. Lett.*, 1998, **286**, 243–252.

8 N. J. DeYonker, B. R. Wilson, A. W. Pierpont, T. R. Cundari and A. K. Wilson, Towards the intrinsic error of the correlation consistent Composite Approach (ccCA), *Mol. Phys.*, 2009, **107**, 1107–1121.

9 T. G. Williams, N. J. DeYonker, B. S. Ho and A. K. Wilson, The correlation Consistent composite Approach: The spin contamination effect on an MP2-based composite methodology, *Chem. Phys. Lett.*, 2011, **504**, 88–94.

10 C. Lee, W. Yang and R. G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, *Phys. Rev. B*, 1988, **37**, 785–789.

11 A. D. Becke, Density-functional exchange-energy approximation with correct asymptotic behavior, *Phys. Rev. A*, 1988, **38**, 3098–3100.

12 A. D. Becke, Density-functional thermochemistry. III. The role of exact exchange, *J. Chem. Phys.*, 1993, **98**, 5648–5652.

13 A. D. Becke, A new mixing of Hartree–Fock and local density-functional theories, *J Chem Phys*, 1993, **98**, 1372–1377.

14 J. P. Perdew, K. Burke and M. Ernzerhof, Generalized gradient approximation made simple, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.

15 M. Ernzerhof and G. E. Scuseria, Assessment of the Perdew-Burke-Ernzerhof exchange-correlation functional, *J. Chem. Phys.*, 1999, **110**, 5029–5036.

16 C. Adamo and V. Barone, Toward reliable density functional methods without adjustable parameters: The PBE0 model, *J Chem Phys*, 1999, **110**, 6158–6170.

17 J. Tao, J. P. Perdew, V. N. Staroverov and G. E. Scuseria, Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids, *Phys. Rev. Lett.*, 2003, **91**, 146401.

18 J. P. Perdew, J. Tao, V. N. Staroverov and G. E. Scuseria, Meta-generalized gradient approximation: Explanation of a realistic nonempirical density functional, *J Chem Phys*, 2004, **120**, 6898–6911.

19 S. Grimme, Accurate Calculation of the Heats of Formation for Large Main Group Compounds with Spin-Component Scaled MP2 Methods, *J Phys Chem A*, 2005, **109**, 3067–3077.

20 Y. Zhao and D. G. Truhlar, A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions, *J Chem Phys*, 2006, **125**, 194101.

21 Y. Zhao and D. G. Truhlar, Density functionals with broad applicability in chemistry, *Acc Chem Res*, 2008, **41**, 157–167.

22 J. D. Chai and M. Head-Gordon, Systematic optimization of long-range corrected hybrid density functionals, *J Chem Phys*, 2008, **128**, 84106.

23 Y. S. Lin, G. D. Li, S. P. Mao and J. D. Chai, Long-Range Corrected Hybrid Density Functionals with Improved Dispersion Corrections, *J Chem Theory Comput*, 2013, **9**, 263–272.

24 W. Zhang, D. G. Truhlar and M. Tang, Tests of Exchange-Correlation Functional Approximations Against Reliable Experimental Data for Average Bond Energies of 3d Transition Metal Compounds, *J Chem Theory Comput*, 2013, **9**, 3965–3977.

25 N. Mardirossian and M. Head-Gordon, ωB97X-V: a 10-parameter, range-separated hybrid, generalized gradient approximation density functional with nonlocal correlation, designed by a survival-of-the-fittest strategy, *Phys Chem Chem Phys*, 2014, **16**, 9904–9924.

26 P. Patel, T. R. L. Melin, S. C. North and A. K. Wilson, *Ab initio composite methodologies: Their significance for the chemistry community*, 2021, vol. 17.

27 P. Patel, D. M. Kuntz, M. R. Jones, B. R. Brooks and A. K. Wilson, SAMPL6 logP challenge: machine learning and quantum mechanical approaches, *J Comput Aided Mol Des*, , DOI:10.1007/s10822-020-00287-0.

28 P. Patel and A. K. Wilson, Domain-based local pair natural orbital methods within the correlation consistent composite approach, *J Comput Chem*, , DOI:10.1002/jcc.26129.

29 A. G. Riojas and A. K. Wilson, Solv-ccCA: Implicit solvation and the correlation consistent composite approach for the determination of pKa, *J. Chem. Theory Comput.*, 2014, **10**, 1500–1510.

30 M. Lundborg, J. Lidmar and B. Hess, The accelerated weight histogram method for alchemical free energy calculations, *J Chem Phys*, , DOI:10.1063/5.0044352.