Benchmarking computational methods to calculate the octanol/water partition coefficients of a diverse set of organic molecules

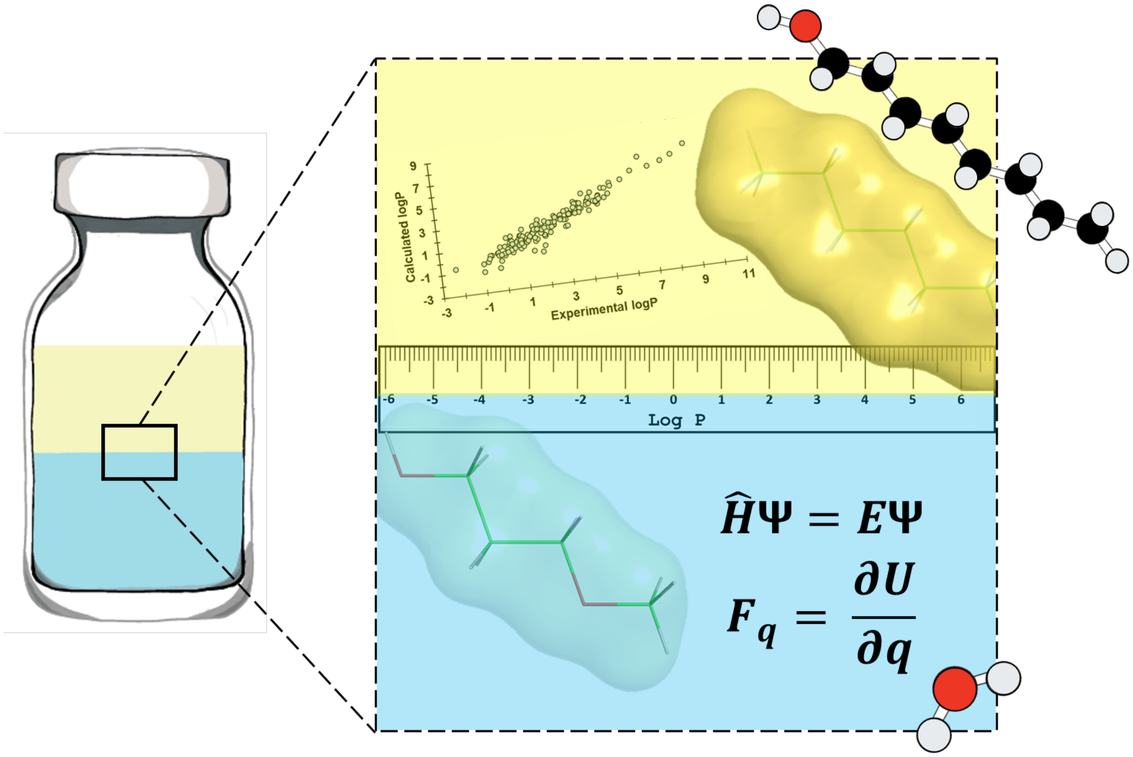
Alondra López-Colón[#], Mariela E Santiago-Mercado[#], Jonathan I. Aguirre-Santiago, Ariana de Jesús-Hernández, Jenlyan Negrón-Hernández, Luis M. Negrón and Dalvin D. Méndez-Hernández*

Department of Chemistry, University of Puerto Rico at Cayey, Cayey, PR 00736

*Corresponding author

[#]These authors contributed equally to the work.

TOC figure:



Abstract

In the discovery process of new drugs and the development of novel therapies in medicine, computational modeling is a complementary tool for the design of new molecules by predicting for example their solubility in different solvents. Here, we benchmarked several computational methods to calculate the partition coefficients of a diverse set of 161 organic molecules with experimental logP values obtained from the literature. In general, density functional theory methods yielded the best correlations and lower average deviations. Although results are obtained faster with semiempirical and molecular mechanics methodologies, these methods yielded higher average deviations and lower correlation coefficients than hybrid density functional theory methods. We recommend the use of an empirical formula to correct the calculated values with each methodology tested.

Keywords

logP, partition coefficient, octanol, water, density functional theory, semiempirical, molecular mechanics

Introduction

Over the past decades, an understanding of the lipophilicity of organic molecules has been essential in designing and synthesizing water-soluble drugs. Moreover, once the solvent solubility is determined, a correlation of the pharmacokinetics with the molecular structure needs to be established for the new drug submitted for clinical trials.[1] Besides the development of profiling techniques based on the Lipinski Rule[2] used in drug discovery[3] , determination of the partition coefficient (P), reported as the base 10 logarithm of P (logP, see equation 1), has been a convenient strategy to provide information about a molecule's lipophilicity. The partition coefficient (P) of a substance is defined as the ratio of its concentration in organic and aqueous phases when the system is in equilibrium.[4,5]

$$\text{logP} = log_{10}\left(\frac{[A]_{organic}}{[A]_{aqueous}}\right) \; where\ [A]_X\ is\ the\ concentration\ of\ A\ in\ X\ phase. (1)$$

For positive logP values, the substance is mainly soluble in the organic phase, while negative values mean that the substance is primarily soluble in the aqueous phase. Experimental determination of logP values can be achieved by various techniques such as shake flask[6], chromatography[7], among others.[8–10] Theoretical estimation of logP values is a common practice to predict the solubility of a molecule during the design phase because it saves valuable materials and reduces time-consuming procedures by targeting compounds with desired solubilities and disregarding those without them. This cost-effective strategy has a significant contribution to the synthesis and development of potential drug candidates.[11] However, the application of the theoretical estimation of logP values is limited to its correlation with the logP values obtained experimentally. The most common method for calculating logP values is with software based on the quantitative structure property relationships (QSPR) model.[1] Nevertheless, the estimated value using the QSPR model has a poor correlation with the experimental logP value because it is limited to a database of experimental logP values of similar molecules.[12] For that reason, computational physics-based free energy models where the solvent is included implicitly with a continuum solvent model are preferred.[13,14]

Recently, Nedyalkova et al. used Density Functional Theory (DFT) to predict the n-octanol/water logP values of 55 organic molecules.[13] They reported the best correlations with the experimental data using the functionals M11 and M06-2X. In this work, we have expanded the list of compounds to a total of 161 molecules by adding 106 molecules from Spafiu et al.[15] to the test set used by Nedyalkova et al.[13] We proceeded to performed calculations with DFT, semiempirical and molecular mechanics (MM) computational methods to benchmark these methodologies. We found the best correlation to be obtained with wB97XD/6-311+G(2d,p).

Methodology:

I. Test set selection

A total of 161 molecules with a variety of functional groups and experimental n-octanol/water logP values were selected from the literature by combining a test set from Neyalkova et al.[13] and another test set from Spafiu et al[15] If the experimental logP value for a compound was reported in both data sets, the average of the two values was used to build the correlations. Figure 1 shows an example molecule for each of the eight molecular families of compounds studied in this work. A complete list of all the molecules with their experimental logP values can be found in the supporting information (SI). The initial structures were built with Avogadro[16] and preoptimized with the Universal Force Field.[17] A complete list of all the molecules in our test set with their experimental and calculated logP values can be found in the SI. Calculations of logP were done for each molecule (Table S1) with each computational methodology discussed below. A summary of the results can be found in Table 1. Scatter plots of correlations 1, 6, 7 and 11 are shown in Figure 2.

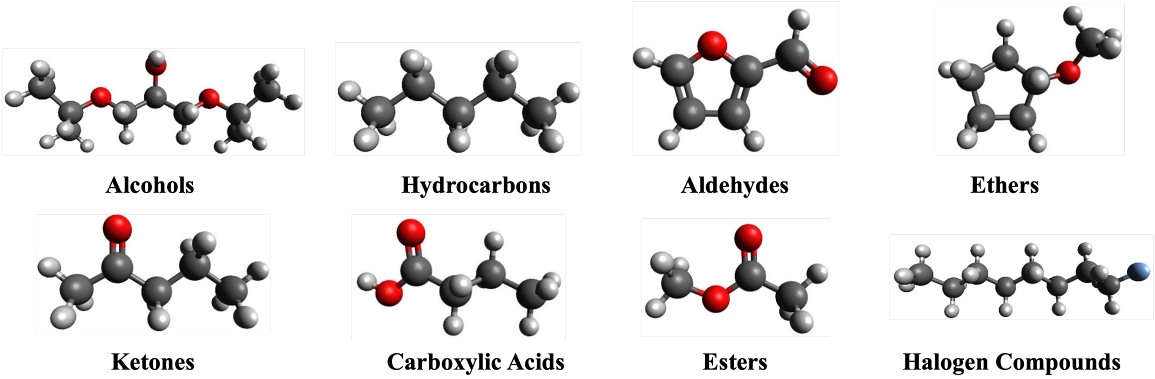


Figure 1. Examples of some of the molecules included in this study. The complete list can be found in the supplementary information (SI).

II. LogP determination with DFT and semiempirical methods

All geometry optimizations and frequency calculations were performed with Gaussian 09[18] using the solvent based on density (SMD) implicit solvent model for water or n-octanol. SMD has been reported to yield the best results in comparison to other continuum based solvent models.[14] The DFT hybrid methods benchmarked with this test set were B3LYP, M062X, and wB97xD functionals with either the 6-31G(d) or 6-311+G(2d,p) as basis sets. The semiempirical methods benchmarked were AM1, PM3, and PM6.

To calculate the logP values for each molecule, we used the Gibbs free energy of the molecules optimized in water and subtracted it from the Gibbs free energy of the molecules optimized in n-octanol ($\Delta G_{o/w}$). Then, logP was calculated according to:

$$logP = \frac{-\Delta G_{o/w}}{2.303RT} \; (2)$$

III. LogP determination with MM methods

Molecular mechanics (MM) methods were carried out using Schrodinger's Maestro 12.5 software and MacroModel tools. The force fields benchmarked were OPLS3e, OPLS, and OPLS2005. The 161 molecules were re-optimized with each of the tested force fields in the gas phase. The OPLS potential parameters used were the dielectric constant as the electric treatment, and charges from the force field. Extended non-bond cutoffs of 8.0, 20.0, and 4.0 Å were used as the parameters for van der Waals, electrostatics, and H-bond, respectively.[19] After re-optimization, the potential parameters were adjusted using n-octanol as the primary solvent, whereas the comparison parameters were activated using the "logP estimation" tool. Water was set as the secondary solvent, to determine the partition coefficients.

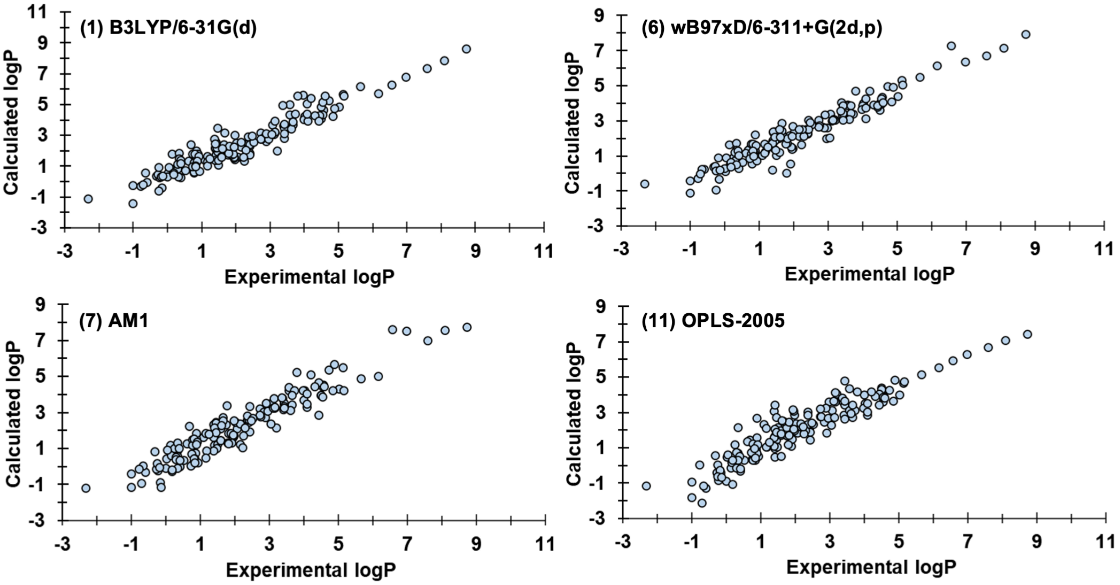


Figure 2. Scatter plots of the calculated vs. experimental logP values for the methodologies 1, 6, 7, and 11.

Results and Discussion

Table 1. Results of the correlations for all methodologies benchmarked.

Correlation number	Method	R^2	MAD	Max residual	Slope	Intercept
1	**B3LYP/6-31G(d)**	0.91	0.50	1.99	0.9970	-0.3061
2	**M062X/6-31G(d)**	0.90	0.49	2.47	0.9763	-0.1987
3	**wB97xD/6-31G(d)**	0.88	0.53	2.04	1.0453	-0.2909
4	**B3LYP/6-311+G(2d,p)**	0.90	0.50	2.58	1.0464	-0.3200
5	**M062X/6-311+G(2d,p)**	0.91	0.48	2.27	0.9926	-0.2305
6	**wB97xD/6-311+G(2d,p)**	0.91	0.47	1.79	1.0886	-0.3424
7	**AM1**	0.89	0.50	2.97	0.9527	0.0338
8	**PM3**	0.85	0.60	2.25	0.9661	-0.1639
9	**PM6**	0.86	0.84	3.30	0.8481	0.9692
10	**OPLS**	0.83	0.67	2.18	0.9527	0.4143
11	**OPLS2005**	0.87	0.54	2.01	0.9894	0.0777
12	**OPLS3e**	0.81	1.44	5.74	0.9186	1.1665

As shown in Table 1, the correlations built with hybrid DFT methods (correlations 1-6) have correlation coefficient (R^2) between 0.87 and 0.91 and mean absolute deviations (MAD) between 0.47 and 0.53. Thus, almost no difference was observed between the functionals and basis sets tested in this work. A possible explanation for this similarity is that the experimental challenges and uncertainties of measuring logP values[7,20] is what limits the strength of the correlations. Correlation 6, (wB97xD/6-311+G(2d,p)) yielded slightly better results than the other DFT correlations (1-5) and the best results overall. This can be rationalized in the light that DFT methods are based on quantum mechanics and have been reported to yield energy calculations with good physical accuracy.[21]

On the other hand, the correlations built with semiempirical methods (correlations 7-9) resulted in lower R^2 values and higher MAD than the hybrid DFT correlations. These results are explicable since semiempirical methods are parametrized with experimental data resulting in approximations that may lead to non-systematic errors. Nevertheless, the

R^2 are all greater than 0.84, which is reasonable considering the significantly faster convergence of the calculations.

Finally, the correlations built with MM methods showed a large variety between them. Correlation 10 showed R^2 values, MAD and Max residuals that are comparable to semiempirical correlations (7-9), while correlation 12 showed MAD and Max residuals that are twice as large. In contrast correlation 11 showed results that were comparable to correlation 3 (a DFT method). These MM calculations run in seconds which significantly reduces the computational resources needed and may be useful as a first approximation. The reason for these faster calculations is that force fields are parametrized based on the Gibbs free energies determined after optimization of each molecule.[22,23] These correlations could be improved by identifying a better extended parametrization for the Maestro's cutoff potential options that considers the van der Waals, electrostatics, and hydrogen bonding interactions which have been well studies and reported in the literature.[19,23] Moreover, since each force field has a particular training set, the nature of the training set must be considered. For example, OPLS and OPLS3e are force fields preferred for biopolymers and carbohydrates, whereas OPLS2005 is the preferred force field for biological systems and organic molecules.[24] This could explain the higher correlation coefficient obtained with OPLS2005.

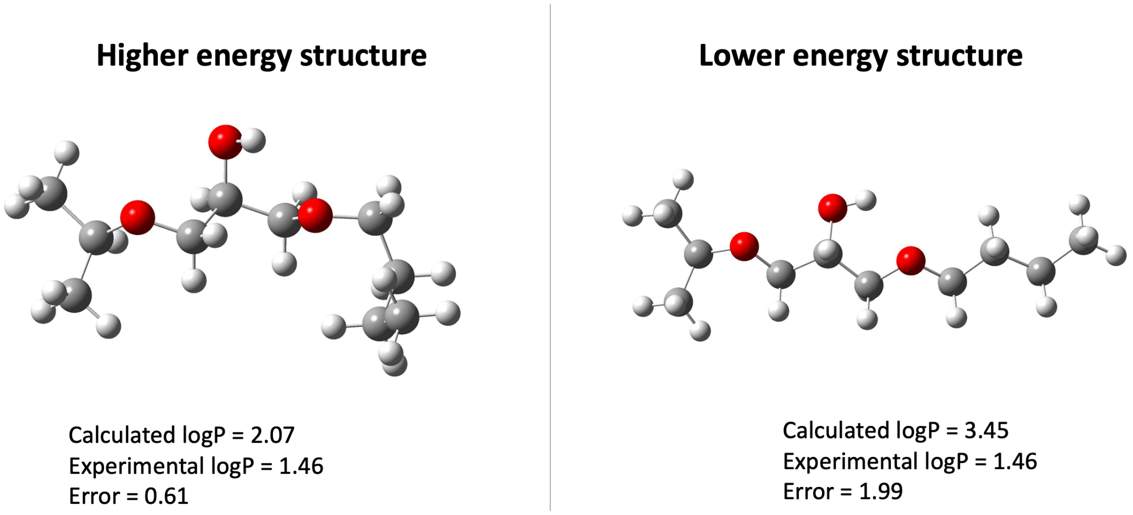


Figure 3. Two conformers of 1-n-butoxy-3-iso-propoxy-2-propanol (molecule 7) optimized using B3LYP/6-31G(d). This is an example of a case in which higher energy structures (left) yield better results than lower energy structures (right).

In the context of conformational structures, using the lowest energy structures was assumed to be crucial to obtain the best computational results as it has been reported by Ho et al.[14] Nevertheless, we found a couple of cases where using higher energy structures yielded better agreement between the calculated and experimental values. One example is shown in Figure 3, where using the higher energy structures (left) resulted in a lower deviation from the experimental logP than using the lower energy structures (right). These results suggest the possibility of improving the correlations by using a Boltzmann population distribution weighted approach[25] to calculate logP values.

Conclusions:

Computational chemistry is a key tool to study the solubility of molecules in complex systems. Therefore, it is important to benchmark the accuracy of computational methods to enable direct comparison to experimental results, and prediction of the lipophilicity and other properties of new compounds.

We tested the accuracy of three DFT functionals, three semiempirical methods, and three molecular mechanics models. Among the methods benchmarked, the most efficient was wB97xD/6-311+G(2d,p). Semiempirical and molecular mechanic methods yielded reasonable correlations, but hybrid DFT methods are recommended for more precise results. To correct systematic errors, we suggest using the correction $\log P_{\text{corrected}} = m\log P_{\text{calculated}} + b$ (where m and b are the slope and intercept found in Table 1, respectively) for calculations done with each methodology presented in Table 1.

These results provide the benchmarking of a diversity of molecules and methodologies that will be useful for future studies in the design and characterization of complex systems such as large molecules, surfactants, and drugs.[2,3,20,26–28] Conducting a Boltzmann population distribution analysis could provide improved calculated logP values as suggested by the results shown in Figure 3 and discussed above. Finally, the inclusion of explicit solvent molecules in combination with implicit solvent models has been reported to be essential to reproduce some experimental results[30–32] and this could further improve the correlations, although some reports claim the contrary.[29]

Acknowledgments:

This project was supported by the PR IDeA Network for Biomedical Excellence PR-INBRE from the National Institute of Health under Award Number 2P20GM103475-18 and the NIH-NIGMS Centers of Biomedical Research Excellence (COBRE) Phase II: Center for Neuroplasticity at the UPR under Award Number 5P20GM103642.

M.E.S.M and A.L.C. were supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R25GM059429-21.

J.N.H. was supported by the National Science Foundation REU under Grant Number 1757365.

Computational resources were provided by the High-Performance Computing Facility at the University of Puerto Rico supported by the Institutional Development Award (IDeA) INBRE Grant Number P20GM103475 from the National Institute of General Medical Sciences (NIGMS), a component of the National Institutes of Health (NIH) and the Bioinformatics Research Core of INBRE. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of NIGMS or NIH.

Author Contributions

D.D.M.H., A.L.C. , M.E.S.M, J.I.A.S., J.N.H. , and A.d.J.H. performed calculations and analyzed the data; D.D.M.H., A.L.C. and L.M.N., designed the study; M.E.S.M., A.L.C. , D.D.M.H., and L.M.N. wrote the paper; and all the authors edited the paper.

Declaration of Interest

The authors declare no competing interests.

References:

1. Hughes, L.D., Palmer, D.S., Nigsch, F., and Mitchell, J.B.O. (2008). Why Are Some Properties More Difficult To Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P. J. Chem. Inf. Model. *48*, 220–232.

2. Chuprina, A., Lukin, O., Demoiseaux, R., Buzko, A., and Shivanyuk, A. (2010). Drug- and Lead-likeness, Target Class, and Molecular Diversity Analysis of 7.9 Million Commercially Available Organic Compounds Provided by 29 Suppliers. J. Chem. Inf. Model. *50*, 470–479.

3. Shultz, M.D. (2019). Two Decades under the Influence of the Rule of Five and the Changing Properties of Approved Oral Drugs. J. Med. Chem. *62*, 1701–1714.

4. Ghose, A.K., Viswanadhan, V.N., and Wendoloski, J.J. (1998). Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragmental Methods: An Analysis of ALOGP and CLOGP Methods. J. Phys. Chem. A *102*, 3762–3772.

5. Bannan, C.C., Calabró, G., Kyu, D.Y., and Mobley, D.L. (2016). Calculating Partition Coefficients of Small Molecules in Octanol/Water and Cyclohexane/Water. J. Chem. Theory Comput. *12*, 4015–4024.

6. Paschke, A., Neitzel, P.L., Walther, W., and Schüürmann, G. (2004). Octanol/Water Partition Coefficient of Selected Herbicides: Determination Using Shake-Flask Method and Reversed-Phase High-Performance Liquid Chromatography. J. Chem. Eng. Data *49*, 1639–1642.

7. Vraka, C., Nics, L., Wagner, K.-H., Hacker, M., Wadsak, W., and Mitterhauser, M. (2017). LogP, a yesterday's value? Nuclear Medicine and Biology *50*, 1–10.

8. Wiczling, P., Kawczak, P., Nasal, A., and Kaliszan, R. (2006). Simultaneous Determination of pKa and Lipophilicity by Gradient RP HPLC. Anal. Chem. *78*, 239–249.

9. Schönsee, C.D., and Bucheli, T.D. (2020). Experimental Determination of Octanol–Water Partition Coefficients of Selected Natural Toxins. J. Chem. Eng. Data *65*, 1946–1953.

10. Henchoz, Y., Guillarme, D., Rudaz, S., Veuthey, J.-L., and Carrupt, P.-A. (2008). High-Throughput log P Determination by Ultraperformance Liquid Chromatography: A Convenient Tool for Medicinal Chemists. J. Med. Chem. *51*, 396–399.

11. Fındık, B.K., Haslak, Z.P., Arslan, E., and Aviyente, V. (2021). SAMPL7 blind challenge: quantum–mechanical prediction of partition coefficients and acid dissociation constants for small drug-like molecules. J Comput Aided Mol Des *35*, 841–851.

12. Jones, M.R., Brooks, B.R., and Wilson, A.K. (2016). Partition coefficients for the SAMPL5 challenge using transfer free energies. J Comput Aided Mol Des *30*, 1129–1138.

13. Nedyalkova, M.A., Madurga, S., Tobiszewski, M., and Simeonov, V. (2019). Calculating the Partition Coefficients of Organic Solvents in Octanol/Water and Octanol/Air. J. Chem. Inf. Model. *59*, 2257–2263.

14. Kundi, V., and Ho, J. (2019). Predicting Octanol–Water Partition Coefficients: Are Quantum Mechanical Implicit Solvent Models Better than Empirical Fragment-Based Methods? J. Phys. Chem. B *123*, 6810–6822.

15. Spafiu, F., Mischie, A., Ionita, P., Beteringhe, A., Constantinescu, T., and Balaban, A.T. (2009). New alternatives for estimating the octanol/water partition coefficient and water solubility for volatile organic compounds using GLC data (Kovàts retention indices). Arkivoc *2009*, 174–194.

16. Hanwell, M.D., Curtis, D.E., Lonie, D.C., Vandermeersch, T., Zurek, E., and Hutchison, G.R. (2012). Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. Journal of Cheminformatics *4*, 17.

17. Rappe, A.K., Casewit, C.J., Colwell, K.S., Goddard, W.A., and Skiff, W.M. (1992). UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. J. Am. Chem. Soc. *114*, 10024–10035.

18. Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Scalmani, G., Barone, V., Mennucci, B., and Petersson, G.A. (2009). Gaussian 09, revision A. 1. Gaussian Inc. Wallingford CT *27*, 34.

19. Reynolds, C.H. (2002). Estimating Lipophilicity Using the GB/SA Continuum Solvation Model: A Direct Method for Computing Partition Coefficients. ACS Publications. https://pubs.acs.org/doi/pdf/10.1021/ci00026a011.

20. Short, J., Roberts, J., Roberts, D.W., Hodges, G., Gutsell, S., and Ward, R.S. (2010). Practical methods for the measurement of logP for surfactants. Ecotoxicol. Environ. Saf. *73*, 1484–1489.

21. Cramer, C.J. (2004). Essentials of Computational Chemistry: Theories and Models 2nd ed. (Wiley).

22. González, M.A. (2011). Force fields and molecular dynamics simulations. JDN *12*, 169–200.

23. MacroModel Reference Manual 262.

24. MacroModel contains a variety of force fields. Which is best for my purpose? | Schrödinger https://www.schrodinger.com/kb/1266.

25. Chaudhuri, S., Hedström, S., Méndez-Hernández, D.D., Hendrickson, H.P., Jung, K.A., Ho, J., and Batista, V.S. (2017). Electron Transfer Assisted by Vibronic Coupling from Multiple Modes. J. Chem. Theory Comput. *13*, 6000–6009.

26. Takahashi, K., Kunishiro, K., Kasai, M., Miike, T., Kurahashi, K., and Shirahase, H. (2008). Relationships between Lipophilicity and Biological Activities in a Series of Indoline-based Anti-oxidative Acyl-CoA:Cholesterol Acyltransferase (ACAT) Inhibitors. Arzneimittelforschung *58*, 666–672.

27. Saranjam, L., Fuguet, E., Nedyalkova, M., Simeonov, V., Mas, F., and Madurga, S. (2021). Prediction of Partition Coefficients in SDS Micelles by DFT Calculations. Symmetry *13*, 1750.

28. Bhal, S.K., Kassam, K., Peirson, I.G., and Pearl, G.M. (2007). The Rule of Five Revisited: Applying Log D in Place of Log P in Drug-Likeness Filters. Mol. Pharmaceutics *4*, 556–560.

29. Chen, J., Shao, Y., and Ho, J. (2019). Are Explicit Solvent Models More Accurate than Implicit Solvent Models? A Case Study on the Menschutkin Reaction. J. Phys. Chem. A *123*, 5580–5589.

30. Méndez-Hernández, D.D., Baldansuren, A., Kalendra, V., Charles, P., Mark, B., Marshall, W., Molnar, B., Moore, T.A., Lakshmi, K.V., and Moore, A.L. (2020). HYSCORE and DFT Studies of Proton-Coupled Electron Transfer in a Bioinspired Artificial Photosynthetic Reaction Center. iScience *23*, 101366.

31. Megiatto, J.D.Jr., Méndez-Hernández, D.D., Tejeda-Ferrari, M.E., Teillout, A.-L., Llansola-Portolés, M.J., Kodis, G., Poluektov, O.G., Rajh, T., Mujica, V., Groy, T.L., et al. (2014). A bioinspired redox relay that mimics radical interactions of the Tyr–His pairs of photosystem II. Nature Chemistry *6*, 423–428.

32. Ortiz-Rodríguez, J.C., Santana, J.A., and Méndez-Hernández, D.D. (2020). Linear correlation models for the redox potential of organic molecules in aqueous solutions. J Mol Model *26*, 70.