Expanded ensemble predictions of toluene–water partition coefficients in the SAMPL9 LogP challenge

Steven Goold, Robert M. Raddi, and Vincent A. Voelz*

Department of Chemistry, Temple University, Philadelphia, PA 19122, USA

E-mail: voelz@temple.edu

Abstract

The logarithm of the partition coefficient (logP) between water and a nonpolar solvent is useful for characterizing a small molecule's hydrophobicity. For example, the water-octanol logP is often used as a predictor of a drug's lipophilicity and/or membrane permeability, good indicators of its bioavailability. Existing computational predictors of water-octanol logP are generally very accurate due to the wealth of experimental measurements, but may be less so for other non-polar solvents such as toluene. In this work, we participate in a Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) logP challenge to examine the accuracy of a molecular simulation-based absolute free energy approach to predict water-toluene logP in a blind test for sixteen drug-like compounds with acid-base properties. Our simulation workflow used the OpenFF 2.0.0 force field, and an expanded ensemble (EE) method for free energy estimation, which enables efficient parallelization over multiple distributed computing clients for enhanced sampling. The EE method uses Wang-Landau flat-histogram sampling to estimate the free energy of decoupling in each solvent, and can be performed in a single simulation. Our protocol also includes a step to optimize the schedule

of alchemical intermediates in each decoupling. The results show that our EE workflow is able to accurately predict free energies of transfer, achieving an RMSD of 2.26 kcal/mol, and R^2 of 0.80. An examination of outliers suggests that improved force field parameters could achieve better accuracy. Overall, our results suggest that expanded ensemble free energy calculations provide accurate first-principles logP prediction.

The logarithm of the partition coefficient (logP) between water and a nonpolar solvent is a highly useful molecular property in medicinal chemistry and pharmacology. LogP measurements for water/octanol partitioning are commonly used to characterize the lipophilicity of drug-like molecules, which can strongly influence their bioavailability and affinity for their targets. Therefore, there is great interest in developing accurate computational methods to predict logP, either by empirical or physics-based methods.[1]

The Statistical Assessment of Modeling of Proteins and Ligands (SAMPL) series of challenges provides an opportunity for research groups to objectively evaluate various methods, through blind prediction of unpublished logP measurements.[1–3] SAMPL has also hosted blind prediction challenges for host–guest affinities,[4–8] pK_a prediction,[9,10] and distribution coefficients (logD).[11] In each case, blind prediction offers the opportunity to examine the accuracy of state-of-the-art methods and to assess where current methods can be improved.

In the previous SAMPL7 logP assessment, challenge participants predicted the water-octanol partition coefficients of 22 molecules, all of which contained sulfonamide groups.[1] The 33 blind predictions of logP submitted for SAMPL7 were classified as empirical, physical QM (quantum mechanics) or physical MM (molecular mechanics). Although methods that achieved a root-mean-squared error (RMSE) of less than 1.0 logP units were mostly empirical, wide interest remains in testing and improving physics-based methods, as these methods should ideally be able to predict logP values from first principles, even in the absence of empirical training data.

Toward this end, we evaluated the accuracy of a molecular simulation approach using expanded ensemble free energy methods to predict logP values.

An expanded ensemble free energy approach to logP prediction

In our expanded ensemble (EE) free energy method, a double-decoupling approach is used to compute the free energy of transfer from vacuum to solvent, through an alchemical transformation in which the nonbonded interactions (electrostatics and van der Waals) are turned off. Given estimated values of ΔG_{tol} and ΔG_{w}, the solvation free energies in toluene and water, respectively, the toluene partition coefficient is calculated as

$$\log_{10} P_{\mathrm{tol/w}} = -\frac{\Delta G_{\mathrm{tol}} - \Delta G_{\mathrm{w}}}{RT(\ln 10)} \tag{1}$$

Expanded Ensemble method. The EE method used here is described in previous works.[12–14] The key feature of EE is the ability to sample multiple thermodynamic ensembles in a single simulation. A coupling parameter $\lambda_i \in [0, 1]$ is used to define a series of $i = 1, \ldots, N$ thermodynamic ensembles where $\lambda_1 = 0$ defines an ensemble with fully-scaled nonbonded interactions, and $\lambda_N = 1$ defines an ensemble where the nonbonded interactions are scaled to zero. Soft-core potentials are used to avoid numerical singularities.

Throughout an EE simulation, a Markov Chain Monte Carlo (MCMC) procedure is used to accept or reject moves between thermodynamic ensembles defined by λ_i and λ_j. The Wang-Landau flat-histogram method[15,16] is used to adaptively learn the values of constant biases $-f_i$ which, when applied to each thermodynamic ensemble i, results in equal probabilities for $i \to j$ and $j \to i$ transitions. When this is achieved, the free energy of the $\lambda = 0 \to 1$ transformation is estimated as $f_N - f_1$. The EE algorithm is available in the GROMACS simulation package.[17] Recent extensions of the EE method have been proposed that combine replica exchange with expanded ensemble sampling,[18,19] but we do not utilize those approaches here.

Because the EE method estimates free energies using a single simulation replica, it is ideally suited for distributed computing applications with an asynchronous client-server model. Our group has recently leveraged Folding@home[20] to perform massively parallel

virtual screening using EE methods.[13,14] Through this work, several methodological issues with have been identified, which are ongoing challenges to be actively addressed. One issue is that the Wang-Landau flat-histogram method results in "saturation of the error" that can lead to premature convergence for fast learning rates.[16] This issue can be compounded by slow convergence due to high-energy barrier conformational transitions in the molecules being decoupled. To deal with these issues, we have found that convergence times and uncertainties of free energies can be estimated by simulating multiple independent EE trajectories.[12,13]

The other major issue with the EE approach is its sensitivity to the chosen schedule of λ_i values. Poor selection of these values can lead to poor MCMC acceptance rates, which in turn causes the simulation to need more time to converge (and potentially error-saturate). To choose optimal schedules for λ_i, we have devised the *pylambdaopt* algorithm,[13] which uses a preliminary round of sampling to infer λ_i values that maximize transition rates between all neighboring thermodynamic ensembles.

Methods

We performed blind predictions of $\log_{10} P_{\mathrm{tol/w}}$ for the sixteen molecules shown in Figure 1 as part of the SAMPL9 logP challenge. A three-part workflow was implemented to (1) prepare systems, (2) perform expanded ensemble simulations on Folding@home and Temple University high-performance computing (HPC) cluster, and (3) analyze the results. All simulations were preformed using GROMACS 2020.3 or GROMACS 2020.4.[17]

System preparation

Molecular topologies. SMILES strings provided by SAMPL9 were converted to three-dimensional chemical structures using the Openeye toolkit.[22] From these, molecule topologies using the OpenFF 2.0.0 force field[23] were constructed using the Open Force Field Toolkit.[24] Partial charges were assigned using AM1-BCC.[25] For simulations in aqueous solvent, the

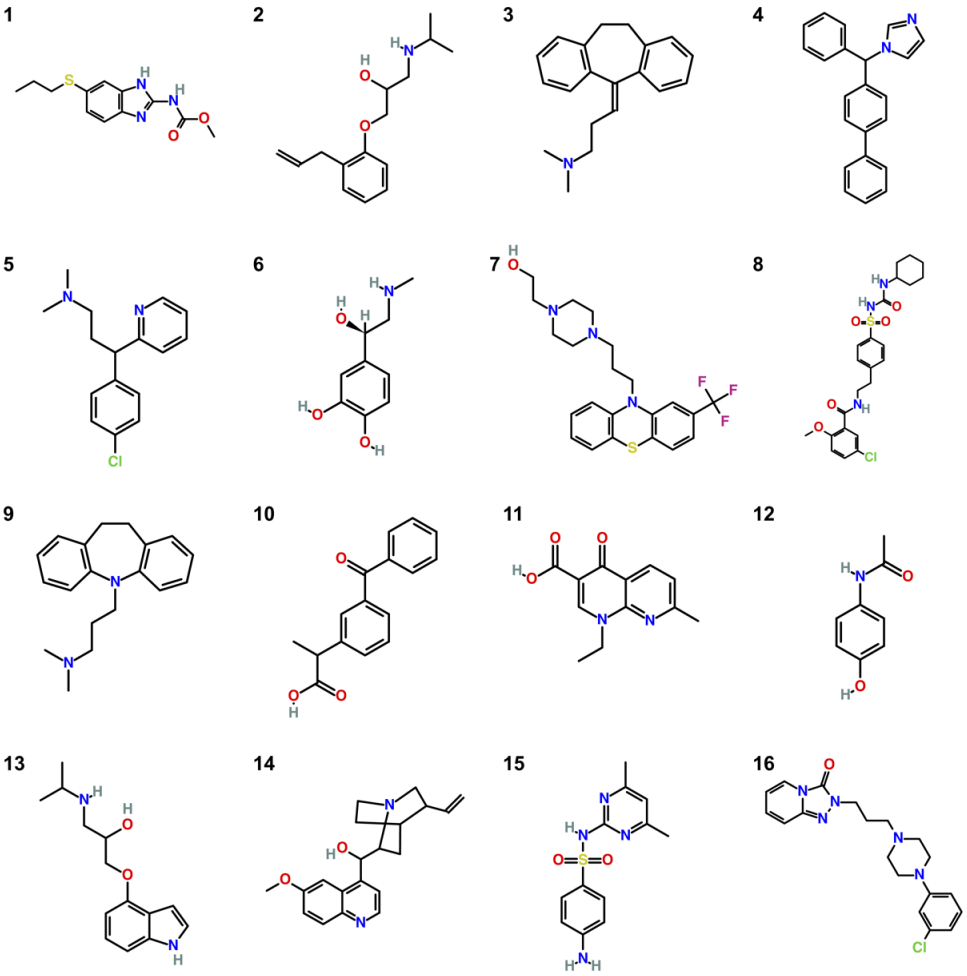

Figure 1: Molecular structures of the molecules in the SAMPL9 logP blind challenge: (1) Albendazole, (2) Alprenolol, (3) Amitriptyline, (4) Bifonazole, (5) Chlorpheniramine, (6) Epinephrine, (7) Fluphenazine, (8) Glyburide, (9) Imipramine, (10) Ketoprofen, (11) Nalidixic acid, (12) Paracetamol, (13) Pindolol, (14) Quinine, (15) Sulfamethazine and (16) Trazodone .[21]

TIP3P water model was used.[26] For simulations in toluene, OpenFF 2.0.0 was used to parameterize the toluene molecule.

Simulation Preparation. An initial solvent box volume of $(6\ \text{nm})^3$ was filled with 7029 water molecules, and an initial solvent box volume of $(3.38\ \text{nm})^3$ was filled with 216 toluene molecules. Solute molecules were inserted into each system and restrained in the center of the box using a harmonic restraint of 1000 kJ nm^{-2}. Systems were energy-minimized using steepest descent before undergoing 200 ps of NVT simulation followed by 200 ps of NPT simulation. All simulations were performed at a temperature of 298.15 K, using velocity Verlet integration with a 2.0-fs time step and with a velocity-rescaling thermostat. The sole exception was for quinine solutes, which used a 1.0-fs time step to avoid instabilities. NPT simulations used Berendsen pressure coupling. These steps were facilitated through in-house scripts utilizing the GromacsWrapper package.[27]

Optimization of lambda values. After NPT equilibration, a short EE simulation was run using a Metropolized-Gibbs MCMC criterion, with attempted moves restricted to nearest-neighbors ($i \rightarrow i \pm 1$). The starting bias was 10 $k_B T$ where k_B is the Boltzmann constant and $T = 298.15$ K is the temperature. Twenty lambda values were used, according to an initial schedule found to work satisfactorily in previous work.[13] Simulation snapshots were saved every 2 ps, with moves to neighboring ensembles attempted every 0.5 ps. Samples of energy differences of snapshots between current and neighboring ensembles ($\Delta u_{i,i-1}$ and $\Delta u_{i,i+1}$, stored in the `dhdl.xvg` output file of GROMACS) were used as input to the *pylambdaopt* algorithm to obtain an optimized schedule of lambda values.

Production simulation

Production simulations were performed in GROMACS on the Folding@Home platform, with 100 independent EE replicas (with different randomized initial velocities) per calculation. All production runs used a velocity Verlet integrator with a 2 fs time step, with the ex-

ception of simulations for quinine, which used a time step of 1 fs , and for which 5 EE replicas were simulated each for 200 ns on the Owlsnest HPC cluster. A velocity-rescaling thermostat was used at a temperature of 298.15 K, and a Berendsen barostat with 2 ps time constant was used at a pressure of 1 bar. Particle-Mesh Ewald electrostatics were used (*pme-order* = 4, *fourierspacing* = 0.10) with nonbonded cutoffs of 0.9 nm and a long-range dispersion correction. Hydrogen bonds were constrained using the LINCS algorithm, and soft-core Lennard-Jones interactions were used (*sc-alpha* = 0.5, *sc-power* = 1, *sc-sigma* = 0.3). Coordinates and energies were saved every 50 ps.

The EE protocol used Wang-Landau flat-histogram sampling[15] with an initial bias increment of $\delta = 10\ k_BT$. When the histogram counts h_i for thermodynamic states i satisfied $\eta \leq h_i/\bar{h} \leq \eta^{-1}$ (where $\eta = 0.7$ and $\bar{h} = (1/N)\sum_{i=1}^{N} h_i$), the bias increment was scaled by 0.8 and all histogram counts were reset to zero. Scaling of the bias increment was discontinued when $\delta < 10^{-5}$.

Analysis of EE simulations

EE simulations were considered to have converged when the bias increment reached a value of 0.02 k_BT. Free energies for each EE replica were estimated as the sample mean of estimates collected after this convergence point. Final estimates of ΔG_{tol} and Δ_{w} and their uncertainties were calculated as the sample means and standard deviations across the EE replicas. Convergence was typically reached within 50–100 ns of simulation, and trajectory lengths typically reached 100–200 ns (Figure 2).

Quantum Mechanical Calculations

To better understand the force field accuracy for fluphenazine, quinine and trazodone, density functional theory (DFT) geometry optimization was performed on simulation snapshots, using the B3LYP functional and cc-DZVP level of theory. Calculations were performed using WebMO[28] with the Gaussian 16 Revision A.03 engine.[29]

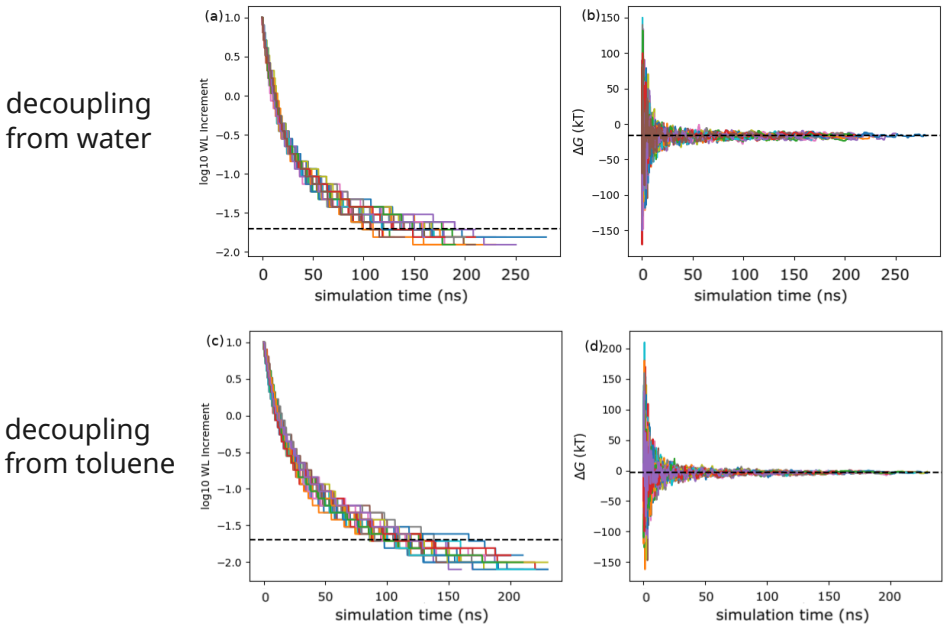

Figure 2: Traces of (a) the Wang-Landau (WL) bias increment over time and (b) the decoupling free energy estimate over time for fifty independent EE trajectories of trazodone in water. Panels (c) and (d) show the corresponding traces for trazodone in toluene.

Results and Discussion

EE accurately predicts toluene–water partition coefficients

Predicted transfer free energies from water to toluene, $\Delta G = \Delta G_{\text{tol}} - \Delta G_{\text{w}}$, show accurate agreement with experimental measurements made by Zamora et al.[30] (Figure 3). Our submitted predictions had a root-mean-squared deviation (RMSD) of 2.26 kcal/mol, a mean signed error (MSE) of 1.09 kcal/mol, a mean unsigned error (MUE) of 1.75 kcal/mol, and a correlation coefficient of $R^2 = 0.80$.

Compared to the other blind predictions submitted by the participants in the SAMPL9 logP challenge, our EE predictions ranked eighth out of 18 total entries by RMSD (Figure 4a). As in the SAMPL7 logP challenge,[1] empirical methods and physics-based QM methods generally outperformed physics-based MM methods, although the top-ranked prediction (RMSD of 1.52 kcal/mol) was from a MM-PBSA method.[31] Of the nine submitted

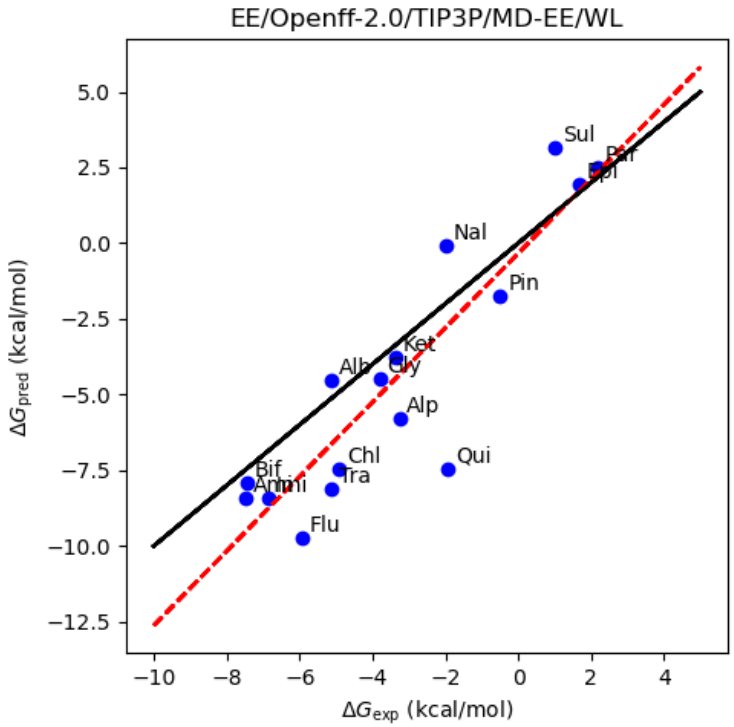

Figure 3: Predicted free energies of transfer from water to toluene, ΔG_{pred}, versus experimental values, ΔG_{exp}, measured by Zamora et al.[30]

predictions from physics-based MM methods, our EE predictions ranked third.

Besides absolute prediction of partition coefficients, we also consider the ability of predictions to correctly rank-order the collection of logP values relative to each other. To quantify this, we used as a statistic the Spearman's rank correlation coefficient, r_s, which is defined as

$$r_s = 1 - \frac{6}{n\left(n^2-1\right)} \sum_{i=1}^{n} d_i^2, \tag{2}$$

where $n = 16$ is the number of ranked items (here, the number of molecules in the SAMPL9 logP challenge) and d_i is the integer difference between the predicted and actual rank.

For all the Physical MM entries submitted, we calculated the value of r_s and compared it with the null distribution, which was computed using 10^5 random permutations of rank orders (Figure 4b). The results show that MM methods with more accurate predictions tend to have more accurate rankings, although interestingly, none of the methods yielded rankings

statistically significant enough to reject the null hypothesis (the smallest p-value calculated was 0.06, for the "MD (OPLS-AA/TIP4P)" submission).

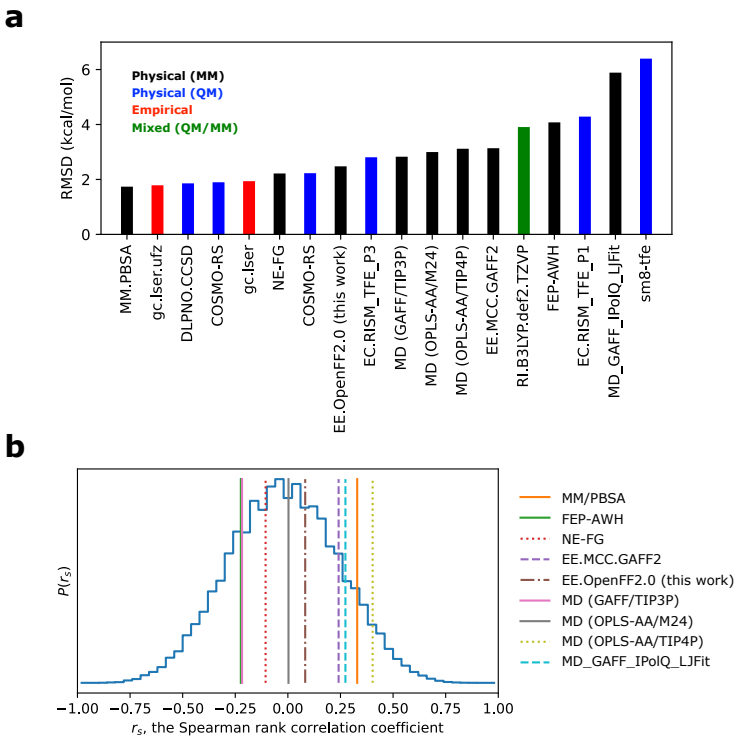

Figure 4: Comparisons of the results of SAMPL9 logP submissions. (a) Prediction accuracies ranked by root-mean-squared deviations from experimental values. Colors denote the kind of method used in each prediction: (black) physical MM, (blue) physical QM, (red) empirical, and (green) mixed QM/MM methods. (b) Spearman's rank correlation coefficients r_s for Physical (MM) methods (vertical lines), plotted in relation to the computed null distribution (blue steps).

An inspection of outliers reveals moderate force field inaccuracy for tertiary amines

An inspection of Figure 3 reveals that the three largest discrepancies between predicted and experimentally measured transfer free energies are for fluphenazine (7), quinine (14), and trazodone (16), with unsigned errors of 5.36, 6.31, and 4.35 kcal/mol, respectively. In all cases, EE predictions underestimate these transfer energies, which means that these molecules are predicted to more favorably partition into toluene than experimentally measurements show.

A commonality shared by these outliers is the presence of non-aromatic nitrogen heterocycles. In contrast, bifonazole (4) and nalidixic acid (11) both contain aromatic nitrogen heterocyclics, and have predicted transfer free energies that are closer to experiment (0.48 and 1.93 kcal/mol unsigned error, respectively).

To examine the extent to which slow conformational sampling could be the cause of large discrepancies for these outliers, we examined the simulated trajectories for these molecules (data not shown). In all cases, dihedral angles incorporating non-aromatic nitrogens show chair-to-chair conversions on the 10–100 ns timescale. This observation, and the similar convergence across all EE simulations (see Figure 2), suggests that the discrepancies are not due to poorly converged sampling from slow conformational dynamics.

Next, we examined the possibility that the outliers might arise from inaccuracies in our chosen force field for tertiary amines. In their article describing the development and performance of the OpenFF Sage 2.0.0 force field, Boothroyd et al. (2023) mention large differences in improper torsion angles between MM- and QM-optimized minima.[23] The largest of these is for the nitrogen-centered improper i4, defined by SMIRKS string `([*:1]~[#7X3](*~[#6X3]):2](~[*:3])~[*:4]")` , although several others (i1, i3, i5) also show deviations. The i4 torsion parameter is difficult to generalize since it covers instances of both planar and pyramidal nitrogens. Of the three outliers, the i4 improper torsion is assigned for one of the nitrogens in fluphenazine, and two of the nitrogens in trazodone; it is not used for either of the quinine nitrogens.

To quantify the extent of nitrogen pyramidalization observed in the simulations of fluphenazine (7), quinine (14), and trazodone (16), we used the Dunitz χ_N parameter, defined as $\chi_N = \omega_1 - \omega_2 + 180°$, where ω_1 and ω_2 are two dihedral angles incorporating the (1,2)- and (1,3)-N-substituents, respectively (Figure 5a). For planar nitrogens, values of χ_N will be near zero, while for a perfectly tetrahedral nitrogen, χ_N values will be near $\pm 60°$.

For fluphenazine (Figure 5b), we found good agreement between the MM minima seen in EE simulations using OpenFF 2.0 and QM minima. The sampled distribution of χ_N for

piperazine nitrogens in the EE simulations were peaked near ±45°, agreeing well with DFT geometry-optimized snapshots from these minima (Figure 5c). The sampled distribution of χ_N for the phenothiazine nitrogen showed planarity; the distribution was centered around zero with a standard deviation around 20°. DFT-optimized snapshots from these simulations indicate QM minima with $\chi_N \approx \pm 12°$ (Figure 5d).

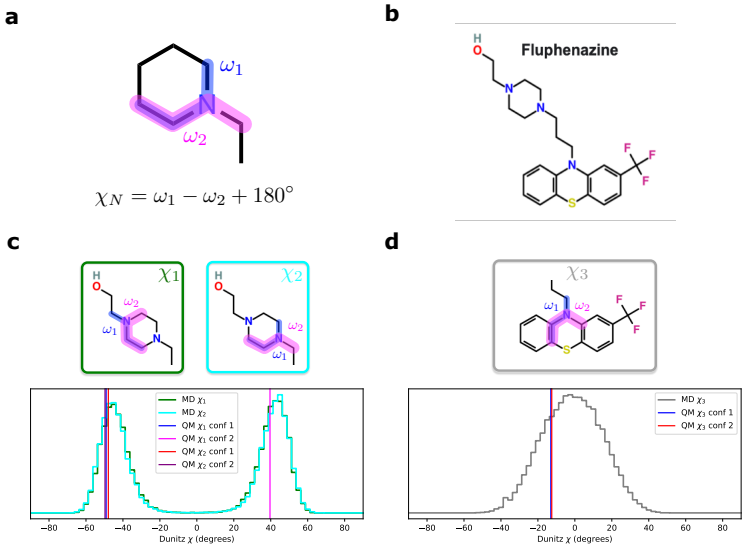

Figure 5: (a) The Dunitz χ_N quantifies the extent of nitrogen pyramidalization using the difference of dihedral angle ω_1 (blue) from dihedral angle ω_2 (magenta). (b) The molecular structure of trazodone. (c) Distributions of Dunitz χ_N parameters χ_1 and χ_2, for the nitrogens in the piperazine group, sampled in expanded ensemble molecular dynamics (MD) simulations of fluphenazine in water. Vertical lines denote the χ_N values of two QM geometry-optimized conformations taken from the simulation. (d) Simulated Dunitz χ_N distributions for the phenothiazine nitrogen, with vertical lines showing values for two QM geometry-optimized conformations.

For the nitrogen in the quinuclidine group of quinine, we found reasonably good agreement between MM and QM minima, which both showing a pyramidal nitrogen (Figure 6a). The distribution of χ_N is narrow (±10°) and peaked around 55°, whereas DFT geometry-optimized snapshots have $\chi_N \approx 64°$.

The greatest disagreement in nitrogen pyramidalization between MM and QM minima was found for piperazine nitrogens in trazodone (Figure 6b). For the nitrogen with fully sp3-hybridized substituents, the sampled distribution of χ_N was peaked near ±45°, agreeing well

with DFT geometry-optimized snapshots with χ_N near $+30°$ and $-38°$. For the nitrogen with the aromatic substituent, simulated χ_N distributions are broadly centered on zero, indicating a planar nitrogen. QM minima, however, suggest χ_N values near $\pm 53°$, suggesting a pyramidal nitrogen.

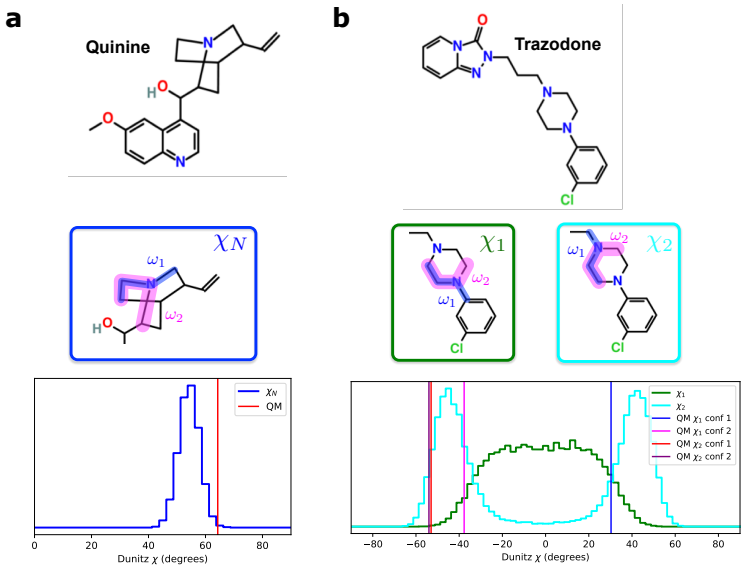

Figure 6: (a) Distributions of Dunitz χ_N parameters χ_1 and χ_2, for the nitrogen in the quinuclidine group of quinine, sampled in expanded ensemble molecular dynamics (MD) simulations of quinine in water. Vertical line denotes the χ_N value of a QM geometry-optimized conformation taken from the simulations. (b) Distributions of Dunitz χ_N parameters χ_1 and χ_2, for piperazine nitrogens of trazodone, sampled in expanded ensemble molecular dynamics (MD) simulations in water. Vertical line denote values of two QM geometry-optimized conformations taken from the simulations.

While these observations are anecdotal, they suggest that improved MM approaches to predicting partition coefficients and other properties may come from improvements in the bonded terms of force fields, specifically torsions. Boothroyd et al. (2023) note that incorrect puckering of small fused heterocycles is of particular concern, as it could lead to "erroneous intramolecular and intermolecular nonbonded interactions, especially in hydrogen bonding interactions and π-stacked configurations".[23] These are exactly the nonbonded interactions that dictate the extent to which molecules partition into nonpolar versus aqueous solvent. In future work, it would be interesting to use our EE approach to compare logP predictions

using general-purpose force fields like OpenFF and GAFF[32] against custom-fit potentials constructed using packages such as OpenFF BespokeFit[33] and AFFDO.[34]

Conclusion

In this work, we have presented the results of an expanded ensemble (EE) free energy method to predict toluene–water partition coefficients in the SAMPL9 logP blind challenge. Our EE method achieved predictions within an RMSD of 2.26 kcal/mol, ranking third out of the nine submissions using physics-based molecular mechanics (MM) methods. Although the most accurate methods for logP prediction continue to be empirical or quantum-mechanical, our inspection of simulated MM versus QM geometries for nitrogen pyramidalization suggests force field improvements may continue to increase the accuracy of physics-based MM methods. The EE method is particularly well-suited for distributed computing platforms, and in the future we expect it to be used more widely for large-scale simulation-based virtual screening.

Acknowledgement

This work is supported by NIH R01GM123296. This research includes calculations carried out on HPC resources supported in part by the National Science Foundation through major research instrumentation grant number 1625061 and by the US Army Research Laboratory under contract number W911NF-16-2-0189. We thank the participants of Folding@home, who made this work possible. We appreciate the National Institutes of Health for its support of the SAMPL project via R01GM124270 to David L. Mobley (UC Irvine).

Data Availability

SAMPL9 logP challenge instructions, experimental data, submissions and analysis are available at `https://github.com/samplchallenges/SAMPL9`. The expanded ensemble (EE) algorithm is implement and freely available in the open-source software package GROMACS (`https://gromacs.org`) Scripts for preparation of EE simulations and data analysis are available at `https://vvoelz.github.io/sampl9-voelzlab`.

References

(1) Bergazin, T. D.; Tielker, N.; Zhang, Y.; Mao, J.; Gunner, M. R.; Francisco, K.; Ballatore, C.; Kast, S. M.; Mobley, D. L. Evaluation of log P, p K a, and log D predictions from the SAMPL7 blind challenge. *Journal of computer-aided molecular design* **2021**, *35*, 771–802.

(2) Bannan, C. C.; Burley, K. H.; Chiu, M.; Shirts, M. R.; Gilson, M. K.; Mobley, D. L. Blind prediction of cyclohexane–water distribution coefficients from the SAMPL5 challenge. *Journal of computer-aided molecular design* **2016**, *30*, 927–944.

(3) Işık, M.; Bergazin, T. D.; Fox, T.; Rizzi, A.; Chodera, J. D.; Mobley, D. L. Assessing the accuracy of octanol–water partition coefficient predictions in the SAMPL6 Part II log P Challenge. *Journal of computer-aided molecular design* **2020**, *34*, 335–370.

(4) Muddana, H. S.; Daniel Varnado, C.; Bielawski, C. W.; Urbach, A. R.; Isaacs, L.; Geballe, M. T.; Gilson, M. K. Blind prediction of host–guest binding affinities: a new SAMPL3 challenge. *26*, 475–487.

(5) Muddana, H. S.; Fenley, A. T.; Mobley, D. L.; Gilson, M. K. The SAMPL4 host–guest blind prediction challenge: an overview. *28*, 305–317.

(6) Yin, J.; Henriksen, N. M.; Slochower, D. R.; Shirts, M. R.; Chiu, M. W.; Mobley, D. L.;

Gilson, M. K. Overview of the SAMPL5 host–guest challenge: Are we doing better? *31*, 1–19.

(7) Rizzi, A.; Murkli, S.; McNeill, J. N.; Yao, W.; Sullivan, M.; Gilson, M. K.; Chiu, M. W.; Isaacs, L.; Gibb, B. C.; Mobley, D. L.; Chodera, J. D. Overview of the SAMPL6 host–guest binding affinity prediction challenge. *32*, 937–963.

(8) Amezcua, M.; Setiadi, J.; Ge, Y.; Mobley, D. L. An overview of the SAMPL8 host–guest binding challenge. *36*, 707–734.

(9) Işık, M.; Rustenburg, A. S.; Rizzi, A.; Gunner, M. R.; Mobley, D. L.; Chodera, J. D. Overview of the SAMPL6 pKa challenge: evaluating small molecule microscopic and macroscopic pKa predictions. *35*, 131–166.

(10) Raddi, R. M.; Voelz, V. A. Stacking Gaussian processes to improve p K a predictions in the SAMPL7 challenge. *Journal of computer-aided molecular design* **2021**, *35*, 953–961.

(11) Bahr, M. N.; Nandkeolyar, A.; Kenna, J. K.; Nevins, N.; Da Vià, L.; Işık, M.; Chodera, J. D.; Mobley, D. L. Automated high throughput pKa and distribution coefficient measurements of pharmaceutical compounds for the SAMPL8 blind prediction challenge. *35*, 1141–1155.

(12) Zhang, S.; Hahn, D. F.; Shirts, M. R.; Voelz, V. A. Expanded Ensemble Methods Can be Used to Accurately Predict Protein-Ligand Relative Binding Free Energies. *Journal of Chemical Theory and Computation* **2021**, *17*, 6536–6547.

(13) Hurley, M. F.; Raddi, R. M.; Pattis, J. G.; Voelz, V. A. Expanded ensemble predictions of absolute binding free energies in the SAMPL9 host–guest challenge. *Physical Chemistry Chemical Physics* **2023**, *25*, 32393–32406.

(14) Novack, D.; Zhang, S.; Voelz, V. A. Massively Parallel Free Energy Calculations for in silico Affinity Maturation of Designed Miniproteins. *bioRxiv* **2024**,

(15) Wang, F.; Landau, D. P. Efficient, multiple-range random walk algorithm to calculate the density of states. *Physical review letters* **2001**, *86*, 2050.

(16) Belardinelli, R. E.; Pereyra, V. D. Wang-Landau algorithm: A theoretical analysis of the saturation of the error. *127*, 184105.

(17) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1*, 19–25.

(18) Hsu, W.-T.; Shirts, M. R. Replica Exchange of Expanded Ensembles: A Generalized Ensemble Approach with Enhanced Flexibility and Parallelizability. *Journal of Chemical Theory and Computation* **2024**, *20*, 6062–6081.

(19) Friedman, A. J.; Hsu, W.-T.; Shirts, M. R. Multiple Topology Replica Exchange of Expanded Ensembles (MT-REXEE) for Multidimensional Alchemical Calculations. 2024; `https://arxiv.org/abs/2408.11038`.

(20) Voelz, V. A.; Pande, V. S.; Bowman, G. R. Folding@ home: Achievements from over 20 years of citizen science herald the exascale era. *Biophysical journal* **2023**, *122*, 2852–2863.

(21) The SAMPL9 Blind Prediction Challenges for Computational Chemistry. `https://github.com/samplchallenges/SAMPL9`, original-date: 2021-08-17T21:01:38Z.

(22) OpenEye Toolkits 2021.1. `http://www.eyesopen.com`.

(23) Boothroyd, S. et al. Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *19*, 3251–3275, Publisher: American Chemical Society.

(24) Wagner, J. et al. openforcefield/openff-toolkit: 0.14.5 Minor feature release. 2023; `https://doi.org/10.5281/zenodo.10103216`.

(25) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *Journal of computational chemistry* **2002**, *23*, 1623–1641.

(26) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics* **1983**, *79*, 926–935.

(27) Beckstein, O. Becksteinlab/GromacsWrapper: release 0.9.0. 2024; `https://zenodo.org/records/11682672`.

(28) Polik, W. F.; Schmidt, J. WebMO: Web-based computational chemistry calculations in education and research. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2022**, *12*, e1554.

(29) Frisch, M. J. et al. Gaussian˜16 Revision A.03. 2016; Gaussian Inc. Wallingford CT.

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

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

(32) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of molecular graphics and modelling* **2006**, *25*, 247–260.

(33) Horton, J. T.; Boothroyd, S.; Wagner, J.; Mitchell, J. A.; Gokey, T.; Dotson, D. L.; Behara, P. K.; Ramaswamy, V. K.; Mackey, M.; Chodera, J. D.; others Open force field BespokeFit: automating bespoke torsion parametrization at scale. *Journal of chemical information and modeling* **2022**, *62*, 5622–5633.

(34) Blanco-Gonzalez, A.; Betancourt, W.; Snyder, R.; Zhang, S.; Giese, T. J.; Goetz, A. W.; Merz, K. M., Jr.; York, D. M.; Aktulga, H. M.; Manathunga, M. Automated Force Field Developer and Optimizer Platform: Torsion Reparameterization. *ChemRxiv* **2024**,