Supporting Information

Sampling of Organic Solutes in Aqueous and Heterogeneous Environments using Oscillating Excess Chemical Potentials in Grand Canonical-Like Monte Carlo-Molecular Dynamics Simulations

Sirish Kaushik Lakkaraju, E Prabhu Raman, Wenbo Yu, Alexander D. MacKerell, Jr.

1Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, 20 Penn St, Baltimore, MD 21201.

* Corresponding Author; email: alex@outerbanks.umaryland.edu
S1. Simulation Details for Aqueous Systems Scheme I and Scheme II
Molecular Dynamics with Periodic Boundary Conditions

These simulations were performed using the CHARMM molecular simulation program. To replicate Scheme I, the required number of molecules to attain 1M of a particular type of solute were randomly placed in a cubic box of 50 Å side. Similarly, to replicate Scheme II, the required number of molecules to attain 0.25 M of each solute type, were randomly inserted in a bulk water cubic box of 50 Å side. The aqueous mixtures were minimized for 5000 steps with the steepest descent algorithm and another 5000 steps of the conjugate gradient algorithm in the presence of periodic boundary conditions. The systems were then heated to 300K at the rate of 5K/ps by periodic reassignment of velocities. Following this, the system was equilibrated for 200 ps using velocity reassignment. The leap frog version of the Verlet integrator with a time step of 1 fs was used for heating and equilibration. Water geometries and covalent bonds involving hydrogen atoms were constrained using the SHAKE algorithm. Long range electrostatic interactions were handled with the particle-mesh Ewald method with a real space cutoff of 12 Å, a switching function was applied to the Lennard-Jones interactions at 10 Å, and a long-range isotropic correction was applied to the pressure for Lennard-Jones interactions beyond the 12 Å cutoff length. Following equilibration, the aqueous mixture boxes were simulated for 15 ns with a time step of 2 fs, at 300K and 1 atm pressure with a Nose-Hoover thermostat, and the Langevin piston barostat.

S2. FragMaps and LGFE scoring

FragMaps
Solute atoms from 10 ps snapshots of the Scheme II aqueous mixture systems in the absence and the presence of the T4-L99A mutant GCMC-MD simulation trajectories were binned into 1 Å x 1 Å x 1 Å cubic volume elements (voxels) of a grid spanning the entire system, and the voxel occupancy for each FragMap atom type was calculated. The normalized distributions were then converted to free energies via a Boltzmann-based transform of the normalized probabilities to yield a grid free energy (GFE) for each fragment type f for the coordinates x,y,z, from which the GFE FragMaps were constructed, as previously reported. All GFE values were capped at 3 kcal/mol. FragMaps were visualized as iso-contour surfaces at a GFE value of -1.2 kcal/mol, unless otherwise noted. The voxel occupancies of the solute atoms in the Tier II aqueous mixture systems are shown in the table below.
**Table S1.** Solute atom occupancies (/ns/M) across the Tier II aqueous systems with finite spherical boundaries.

| Fragment Atoms       | GCMC-MD rad 25 Å |
|----------------------|------------------|
| Benzene carbons      | 0.041            |
| Propane carbons      | 0.016            |
| Acetaldehyde oxygen  | 0.006            |
| Methanol hydrogen    | 0.005            |
| Methanol oxygen      | 0.005            |
| Formamide hydrogen   | 0.018            |
| Formamide oxygen     | 0.009            |
| Acetate oxygens      | 0.016            |
| Methylam. Polar hydrogens | 0.018 |

**LGFE scoring**

Ligand atoms were classified into FragMap types, for which an assignment map was created. The rules for assignment are presented in our previous work. Each classified atom of a ligand with coordinates \((x_i, y_i, z_i)\) was assigned a score equal to the GFE value of the corresponding FragMap type \((f)\), \(GFE^f_{x_i,y_i,z_i}\), of the voxel it occupies. LGFE is then the sum of each of these GFE values for all the classified ligand atoms. LGFE was calculated as Boltzmann weighted averages over two conformational ensembles of the ligands.

**MD conformational ensemble (for LGFE\(^\text{MD}\)):** Initial ligand conformations were extracted from the corresponding co-crystal coordinate PDB files and the automated CGenFF parametrization algorithm\(^{13,14}\) was used to obtain the topology and parameters in the context of CGenFF\(^{15}\). The ligand conformation from the co-crystal structure was extracted and aligned with the protein conformation used in the SILCS simulations. The alignment was done based on optimal alignment of the backbone \(C_\alpha\) atoms in the two protein structures. The complex was minimized using the SD algorithm for 50 steps. This minimized conformation was subject to a 10 ps MD simulation, with snapshots output every 0.2 ps. During the dynamics, all protein atoms farther than 8 Å around any ligand atom were restrained using a force constant of 1 kcal/mol/Å\(^2\). This process was repeated on 40 protein conformations obtained from the GCMC-MD SILCS simulations, equally spaced in time (cycle 10, 20, 30, ...), yielding at total of 1000 ligands conformations for estimation of the Boltzmann weighted average LGFE.

**MC conformational ensemble (for LGFE\(^\text{MC}\)):** Sampling of the ligand was also performed in the “field” of the GFE FragMaps using Metropolis Monte Carlo (MC) steps. An in-house suite of programs was used to setup and run the MC simulations. The ligand had rotational, translational and intra-molecular degrees of freedom. The ligand had no rotational restraints, but its center of mass (CoM) was restrained to lie within 2.5 Å of the CoM of the ligand crystal conformation using a flat bottom restraint. Intramolecular degrees of freedom consisted of rotatable bonds, which were automatically detected based on the CGenFF molecular topology. All acyclic non-terminal bonds were considered rotatable, with the exception of bonds ending in methyl or \(\text{NH}_3^+\) groups. The force-field terms corresponding to the intramolecular degrees of freedom comprised of dihedral, van der Waals (vdW) and electrostatic terms. Due to the absence of protein and solvent during these simulations a distance dependent dielectric \((\approx 4|\mu|)\) was used to evaluate the
intramolecular electrostatic contributions to prevent their overestimation. The energy computed during the Metropolis MC can be written as follows.

\[ E = E_{\text{vdw,intra}} + E_{\text{elec,intra}} + E_{\text{dih,intra}} + \text{LGFE} \]

For each ligand, 20 different MC simulations (each run for 10,000,000 steps; snapshots recorded every 10,000 steps) were run, where for each of them the ligand is initially randomly placed close to the protein active site. An average LGFE is first calculated over each of these 20 MC simulations. LGFE\textsuperscript{MC} is then the Boltzmann weighted average over these LGFE values obtained across the 20 MC simulations.

**S3. Supporting Figures and Tables**

Figure S1. A. Distribution of benzene in a spherical boundary aqueous system. When solute and waters share the same spherical boundary, hydrophobic solutes accumulate at the edge of the wall. B. To avoid this, the system A is immersed in a larger system B as described in the main text.
Figure S2. Distribution of $\mu_{ex}$ values across cycles 150-200 from the Scheme I GCMC-MD simulations with a spherical system of radius 25 Å. The dashed lines indicate the average $\mu_{ex}$ values in Table 1.

Figure S3. Radial distribution functions ($g(r)$) for the solute atoms from the Scheme I and II GCMC-MD and PBC MD simulations. Black: MD with PBC replicating Scheme I, Red: MD with PBC replicating Scheme II, Green: GCMC-MD Scheme I, Blue: GCMC-MD Scheme II. For benzene (BENZ), propane (PRPN), methylammonium (MAMM) and acetate (ACET), $g(r)$ is measured based on the massless particles (LP) that were added to their center of masses to prevent aggregation. For methanol (MEOH), formamide (FORM) and acetaldehyde (AALD), $g(r)$ is measured between the polar hydrogens or oxygens.
Figure S4. Concentration (M) and $\mu_{\text{ex}}$ (kcal/mol), as a function of the number of GCMC-MD cycles from Scheme I and Scheme II GCMC-MD aqueous systems with $\mu_{\text{ex}}$ fixed at HFE (black and green) or fluctuated by $d\mu_{\text{ex}}$ (red and blue) respectively.
Fig S5. $\mu_{ex}$ (kcal/mol), probabilities of insertion+deletion ($P_{ins+del}$) and translation+rotation ($P_{trans+rot}$) as a function of the number of GCMC-MD cycles from Scheme I and Scheme II GCMC-MD of aqueous systems with $\mu_{ex}$ fixed at HFE (black and green) or fluctuated by $d\mu_{ex}$ (red and blue) respectively.
**Figure S6.** Setup for the Scheme I GCMC-MD with the T4-L99A mutant. The GCMC-MD is restricted to the 20 Å radius active sphere (system A) with the center at the active site of the T4-L99A mutant defined by residues Ala 99 and Met 102. The system A is encompassed in the system B which is a PBC box with walls about 12 Å away from the protein surface. System B contains both waters and benzenes at 55 M and 1 M, respectively. Waters and benzenes within the active sphere are colored blue and purple, respectively.

**Table S2.** Average $\mu_{ex}$ of the solute and the water through the Scheme I and Scheme II GCMC-MD simulations with a spherical system of radius 25 Å, obtained from cycles 50-200.

| Fragment   | Scheme I        | Scheme II       |
|------------|-----------------|-----------------|
|            | $\mu_{ex}$(kcal/mol) | Conc (M) | $\mu_{ex}$(kcal/mol) | Conc (M) |
| Benzene    | -0.62±0.47      | 1.21±0.21      | -0.94±0.23      | 0.32±0.21 |
| Propane    | 1.39±0.23       | 1.32±0.13      | 1.31±0.21       | 0.37±0.13 |
| Acetaldehyde | -3.1±0.71    | 1.1±0.81       | -2.92±0.43      | 0.24±0.12 |
| Methanol   | -5.73±0.31      | 1.1±0.34       | -4.92±0.21      | 0.25±0.11 |
| Formamide  | -14.2±1.22      | 1.03±0.43      | -11.51±1.31     | 0.22±0.12 |
| Acetate    | -50.2±0.71      | 0.83±0.31      | -52.1±0.62      | 0.24±0.04 |
| Meth. Amm. | -58.1±0.42      | 0.71±0.32      | -56.1±0.74      | 0.22±0.17 |
| Water      | -5.2±0.12       | 54.4±1.1       | -5.0±0.13       | 55.1±0.41 |
**Table S3.** Average $\mu_{ex}$ of acetaldehyde and methanol through the Scheme I simulation with and without MD at the end of GCMC every iteration in a spherical system of radius 25 Å, obtained from the last 50 cycles.

| Fragment     | Scheme I GCMC-MD | Scheme I GCMC |
|--------------|------------------|---------------|
|              | $\mu_{ex}$(kcal/mol) | Conc (M) | $\mu_{ex}$(kcal/mol) | Conc (M) |
| Acetaldehyde | -3.10±0.54       | 1.01±0.11    | -2.93±0.1     | 1.04±0.22 |
| Methanol     | -5.79±0.23       | 1.30±0.49    | -5.32±0.3     | 1.20±0.32 |

**Figure S7.** Selected GFE Fragmaps at the ligand binding site of the T4-lysozyme L99A from A) 10×50 ns GCMC-MD simulation and B) 100 million steps of GCMC, along with the minimized crystal conformations of the 9 ligands; protein atoms occluding the view of the binding pocket were removed for clear visualization. Without MD, the pocket in the neighborhood of Ser117 and Leu118 does not open up for sampling visualization.

**References**

(1) Brooks, B. R.; Brooks, C. L.; MacKerell, A. D.; Nilsson, L.; Petrella, R.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S. CHARMM: the biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545-1614.
Refinement of protein conformations using a macromolecular energy minimization procedure. *J. Mol. Biol.* **1969**, *46*, 269-279.

Conjugate-direction minimization: an improved method for the refinement of macromolecules. *Acta Crystallogr A* **1992**, *48*, 912-916.

Computer simulation of liquids; Oxford university press, 1989.

Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **1980**, *72*, 2384.

Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of \( n \)-alkanes. *J. Comput. Phys.* **1977**, *23*, 327-341.

Particle mesh Ewald: An \( N \log(N) \) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089.

New spherical-cutoff methods for long-range forces in macromolecular simulation. *J. Comput. Chem.* **2004**, *15*, 667-683.

A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.* **1984**, *52*, 255-268.

Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695.

Molecular-dynamics simulation-the Langevin piston method. *J. Chem. Phys.* **1995**, *103*, 4613-4621.

Inclusion of multiple fragment types in the Site Identification by Ligand Competitive Saturation (SILCS) approach. *J. Chem. Info. Model.* **2013**.

Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. *J. Chem. Info. Model.* **2012**, *52*, 3144-3154.

Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Info. Model.* **2012**, *52*, 3155-3168.

CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **2010**, *31*, 671-690.