Supporting information for: grand: A Python Module for Grand Canonical Water Sampling in OpenMM

Marley L. Samways,[†] Hannah E. Bruce Macdonald,[‡] and Jonathan W. Essex*,[†]

†*School of Chemistry, University of Southampton, Southampton, SO17 1BJ, UK*

‡*Computational and Systems Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA*

E-mail: j.w.essex@soton.ac.uk

Calculation of excess chemical potential and standard volume of bulk water

For internal consistency in the GCMC/MD simulations, it was deemed appropriate to calculate the values of the excess chemical potential and standard state volume of water from simulations, rather than using the experimental values. The former is calculated as the hydration free energy of water, and the latter as the average volume per water molecule. In both cases, the simulations were carried out at constant pressure, starting from a pre-equilibrated cubic box containing 2094 TIP3P water molecules, with an initial density of 0.978 g mL^{-1}. The hydration free energy of water was calculated using the multistate Bennett acceptance ratio (MBAR) method,[S1,S2] with 30 λ-values from 1 to 0, where $\lambda = 1$ indicates a fully interacting water, and $\lambda = 0$ indicates a fully decoupled water, with softcore potentials applied to avoid instabilities.[S3] For each value of λ, 1000 potential energy samples

were collected, with 10 ps of MD carried out between samples, giving a total of 300 ns of simulation. The samples were then processed to remove unequilibrated[S4] and correlated data,[S5] before calculating the free energy, all using the functions provided by version 3.0.1 of the *pymbar* Python module.[S6] The average volume per water molecule was calculated as the mean value of samples collected every 5 ps for 50 ns.

A total of 50 hydration free energy calculations for water were performed, arriving at a mean and associated standard error of -6.087±0.005 kcal mol^{-1}, from which we take $\mu'_{sol} = -6.09$ kcal mol^{-1}. The average volume per water molecule was calculated 10 times, yielding a mean value, with standard error, of 30.3454±0.0006 Å^3 , from which we take $V^\circ = 30.345$ Å^3. In each case, the number of independent repeats was increased until the standard errors obtained were sufficiently low.

Unless explicitly stated otherwise, all B_{equil} values reported in this work were calculated using the values of μ'_{sol} and V° determined from these calculations.

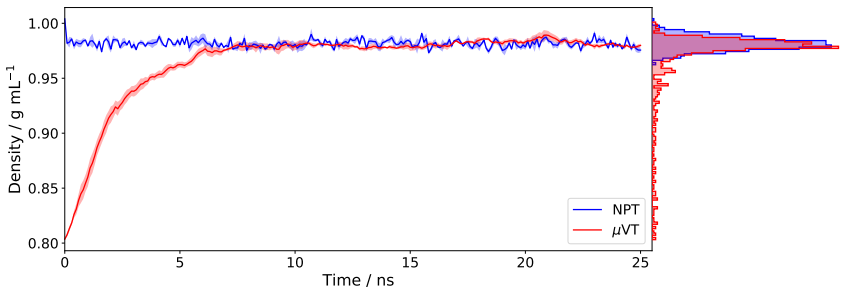
Neglect of Long-Range Dispersion Corrections

As stated in the main text, long-range dispersion corrections were not employed in this work, for reasons of efficiency. This is because, to prevent singularities involving non-interacting waters (those which have either been deleted, or not yet inserted), softcore potentials are used for all interactions involving water molecules. This involves the use of 'CustomNonbondedForce' objects in OpenMM to calculate these interactions, for which the analytical solution to the calculation of long-range dispersion interactions is not implemented, and a numerical solution must be used. Whilst this numerical calculation has been written efficiently to reduce repeated calculations, this is disturbed when waters are inserted or deleted, upsetting these efficiency measures, and making the calculation of the potential energy of each proposed move much more expensive. Therefore, the decision was made to neglect these long-range dispersion effects in this work. We plan to find an efficient solution to this

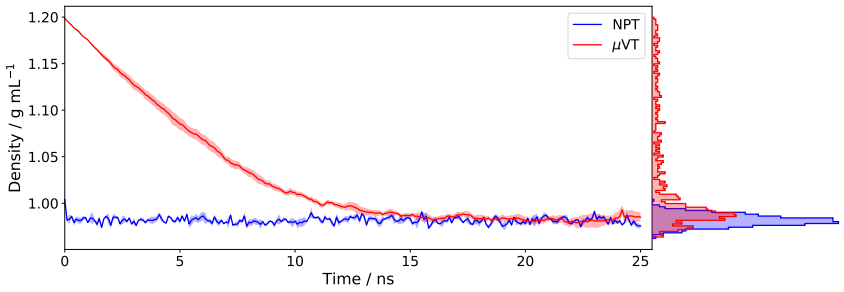
problem in future work.

Further tests involving bulk water

Bulk water was also simulated using initial densities of 0.804 and 1.199 g mL^{-1}, to test the ability of the GCMC/MD implementation to return these systems to the equilibrium densities. These simulations were carried out under the same conditions as those reported in the main text, only with the initial particle number adjusted to modify the densities appropriately. These simulations were run for 25 ns, with densities reported every 0.1 ns, with three independent repeats carried out in each case. An additional NPT run was carried out at this shorter timescale, for comparison, using the original water box with a starting density of 1.004 g mL^{-1}. These results are shown below in Figure S1, where it is shown



(a) Starting density of 0.804 g mL^{-1} for GCMC/MD.

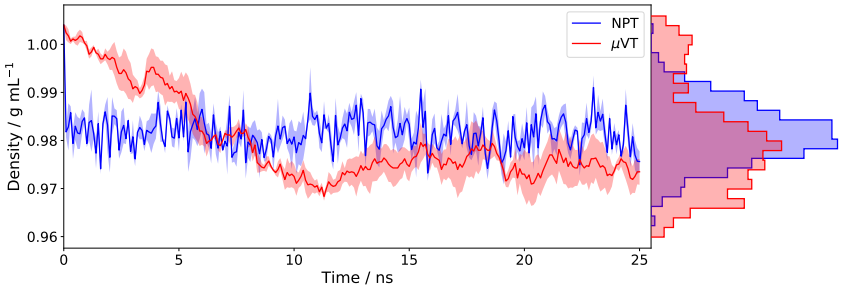


(b) Starting density of 1.199 g mL^{-1} for GCMC/MD.

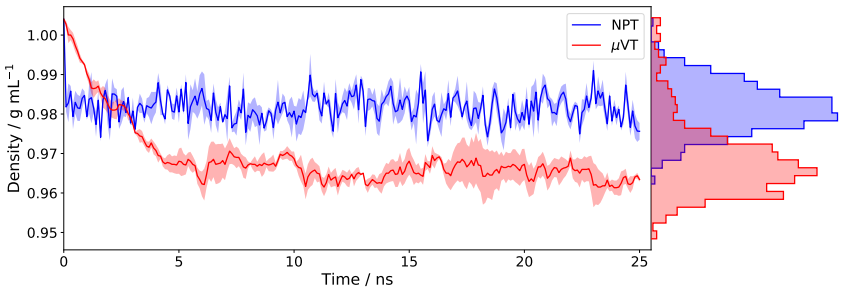
Figure S1: Variation in the water density over time for the GCMC/MD simulations performed (red data), when starting the simulation from water boxes with different densities, compared with the NPT results (blue data). The solid line represents the mean density over the three independent repeats, and the shaded region represents the associated standard error. The histograms were plotted using all density samples.

that GCMC/MD converges these systems to the same mean density as observed in the NPT simulations. Whilst the distributions are not completely equilibrated, owing to the short simulation time (and that GCMC/MD is much less efficient for such large systems), there is no clear drift away from this density.

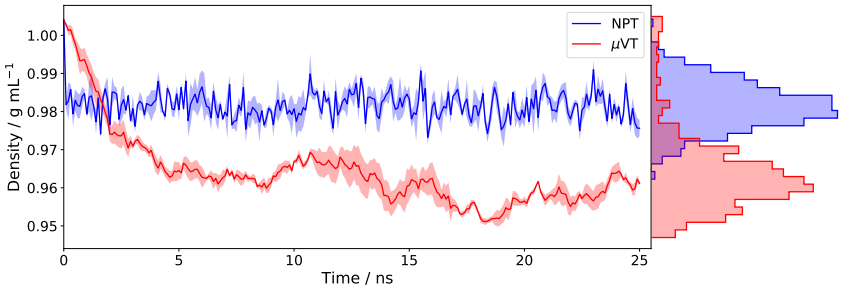
Further, additional short GCMC/MD runs were carried out with an initial density of 1.004 g mL^{-1}, using different values of the excess chemical potential to demonstrate the



(a) $\mu'_{sol} = -6.15$ kcal mol^{-1}.



(b) $\mu'_{sol} = -6.20$ kcal mol^{-1}.



(c) $\mu'_{sol} = -6.25$ kcal mol^{-1}.

Figure S2: Variation in the water density over time for the GCMC/MD simulations performed (red data), using three different values of the excess chemical potential (μ'_{sol}) compared with the NPT results (blue data). The solid line represents the mean density over the three independent repeats, and the shaded region represents the associated standard error. The histograms were plotted using all density samples.

sensitivity of the results to this parameter. Values of -6.15, -6.20 and -6.25 kcal mol^{-1} were tested for the excess chemical potential, with the standard state volume set to 30.345 Å^{-1} in all cases. These results are shown in Figure S2. As can be seen, the results appear very sensitive to correct parametrization of the excess chemical potential, as even a difference of 0.06 kcal mol^{-1} produces a notable difference from the results obtained using an NPT simulation. However, it is important to note that GCMC simulations are not typically applied to systems with such large volumes, so differences will be magnified in this case.

References

(S1) Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *J. Comput. Phys.* **1976**, *22*, 245–268.

(S2) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129*, 124105.

(S3) Beutler, T. C.; Mark, A. E.; van Schaik, R. C.; Gerber, P. R.; van Gunsteren, W. F. Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem. Phys. Lett.* **1994**, *222*, 529–539.

(S4) Chodera, J. D. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *J. Chem. Theory Comput.* **2016**, *12*, 1799–1805.

(S5) Chodera, J. D.; Swope, W. C.; Pitera, J. W.; Seok, C.; Dill, K. A. Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations. *J. Chem. Theory Comput.* **2007**, *3*, 26–41.

(S6) Beauchamp, K. A.; Chodera, J. D.; Naden, L.; Shirts, M. R.; Martiniani, S.; Stern, C. D.; McGibbon, R. T.; Gowers, R.; Barnett, J. pymbar. 2017; `https://github.com/choderalab/pymbar`, (Date accessed: March 3, 2020).