Supporting Information: The role of force field parameter uncertainty in the prediction of the pressure-viscosity coefficient

Richard A. Messerly

Thermodynamics Research Center, National Institute of Standards and Technology, Boulder, Colorado, 80305

Michelle C. Anderson

Thermodynamics Research Center, National Institute of Standards and Technology, Boulder, Colorado, 80305

S. Mostafa Razavi

Department of Chemical and Biomolecular Engineering, The University of Akron, Akron, Ohio, 44325-3906

J. Richard Elliott

Department of Chemical and Biomolecular Engineering, The University of Akron, Akron, Ohio, 44325-3906

SI.I. Input files

We provide example input files for simulating 2,2,4-trimethylhexane at 293 K with the Potoff force field in GROMACS (see attached .gro, .top, and .mdp files). Additionally, all files necessary to generate the results from this study can be found at www.github.com/ramess101/IFPSC_10.

Email addresses: richard.messerly@nist.gov (Richard A. Messerly), michelle.anderson@nist.gov (Michelle C. Anderson), sr87@zips.uakron.edu (S. Mostafa Razavi), elliot1@uakron.edu (J. Richard Elliott)

SI.II. MCMC from scoring function

Markov Chain Monte Carlo (MCMC) requires an expression for the likelihood function (L) and, in particular, the log (or natural log) of the likelihood function ($\ln(L)$). For example, by assuming a uniform prior in the model parameters and a symmetric proposal distribution, the Metropolis-Hastings acceptance criterion for an MCMC move only depends on L such that

$$\alpha = \frac{L(D|\theta_{\text{new}})}{L(D|\theta_{\text{old}})} \tag{1}$$

where D are the data, θ are the model parameters, and α is the probability of accepting a proposed or "new" parameter set (θ_{new}) given a previous or "old" parameter set (θ_{old}). For computational reasons, it is common to perform MCMC using the log-likelihood such that

$$\ln(\alpha) = \ln(L(D|\theta_{\text{new}})) - \ln(L(D|\theta_{\text{old}})) \tag{2}$$

By contrast, Mick et al. optimize the Potoff CH and C parameters using a scoring function (S) that weights the deviations for several different properties and their derivatives. This section describes how we translate the scoring function into a log-likelihood function to enable direct application of MCMC.

SI.II.1. Derivation

Standard least squares minimization is mathematically equivalent to maximizing the likelihood function when assuming the errors follow a normal distribution. This can be readily verified from the following expression

$$\begin{aligned} L(D|\theta) = \prod_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-1}{2\sigma^2}(y(\theta) - D_i)^2\right] = \frac{1}{\sqrt{2\pi^n\sigma^{2n}}} \exp\left[\frac{-1}{2\sigma^2}\left(\sum_i (y(\theta) - D_i)^2\right)\right] \\ = \frac{1}{\sqrt{2\pi^n\sigma^{2n}}} \exp\left(\frac{-SSE(\theta)}{2\sigma^2}\right) \end{aligned} \tag{3}$$

where n is the number of data points, σ is the standard deviation (which is assumed to be equal for all data points), $y(\theta)$ is the model estimate, and $\sum_i (y(\theta) - D_i)^2$ is the sum-squared-error (SSE).

The log-likelihood can then be expressed as

$$\ln(L(D|\theta)) = \ln\left(\frac{1}{\sqrt{2\pi^n\sigma^{2n}}}\right) - \frac{SSE(\theta)}{2\sigma^2} \tag{4}$$

Substitution of Equation 4 into Equation 2 yields the MCMC acceptance probability

$$\ln(\alpha) = \frac{SSE(\theta_{\text{old}}) - SSE(\theta_{\text{new}})}{2\sigma^2} \tag{5}$$

note that the order of "new" and "old" changes due to the negative sign in Equation 4. Also, note the cancellation of the term in Equation 4 that does not depend on θ.

To this point, we have assumed that the standard deviation is constant with respect to the data. Relaxing this assumption, the likelihood is expressed as

$$\begin{aligned} L(D|\theta) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[\frac{-1}{2\sigma_i^2}(y(\theta) - D_i)^2\right] &= C\exp\left[\sum_i \frac{-1}{2\sigma_i^2}(y(\theta) - D_i)^2\right] \\ &= C\exp\left[\sum_i -w_i(y(\theta) - D_i)^2\right] = C\exp\left(-WSSE(\theta)\right) \end{aligned} \tag{6}$$

where σ_i varies for D_i, C is a normalization constant equal to $\prod_i (2\pi\sigma_i^2)^{-1}$, and $WSSE$ is the weighted-sum-squared-error with weights (w_i) equal to σ_i^{-2}. The log-likelihood can then be expressed as

$$\ln(L(D|\theta)) = \ln(C) - WSSE(\theta) \tag{7}$$

and substitution into Equation 2 yields the MCMC acceptance probability for the weighted-sum-squared-error

$$\ln(\alpha) = WSSE(\theta_{\text{old}}) - WSSE(\theta_{\text{new}}) \tag{8}$$

Potoff's scoring function (S) is not simply the sum-squared-error or even the weighted-sum-squared-error. S is expressed in terms of the absolute percent deviation and different properties are assigned different weights that are not necessarily the inverse variance (σ_i^{-2}). Therefore, minimizing S is not equivalent to maximizing the likelihood of a normal distribution for neither constant nor varying σ. However, we can still apply the maximum likelihood criterion for Potoff's

scoring function such that

$$L(D|\theta) = \prod_i C_i \exp\left(w_i \left|\frac{y(\theta) - D_i}{D_i}\right|\right) = C \exp\left(\sum_i w_i \left|\frac{y(\theta) - D_i}{D_i}\right|\right) = C \exp\left(-S(\theta)\right) \quad (9)$$

where C_i and C are normalization constants, and w_i are the weights assigned by Potoff (e.g., 0.6135 for saturated liquid density, ρ_{liq}). Note that although Equation 9 utilizes an absolute percent deviation, rather than a weighted-sum-squared-error, the final expression is still analogous to Equation 6. Therefore, the log-likelihood for Potoff's scoring function is

$$\ln(L(D|\theta)) = \ln(C) - S(\theta) \quad (10)$$

and substitution into Equation 2 yields the MCMC acceptance probability for Potoff's scoring function

$$\ln(\alpha) = S(\theta_{\text{old}}) - S(\theta_{\text{new}}) \quad (11)$$

SI.II.2. Implementation

With Equation 11, all that remains to perform MCMC is a way to compute S for any θ. This is achieved by smoothing/interpolating the raw scoring function values (utilizing SciPy's RectBivariateSpline function in the interpolation sub-package) over the two-dimensional grids of ϵ_{CH}-σ_{CH} and ϵ_{C}-σ_{C}. The values of S were obtained through private communication with Potoff's group. Only the scoring function values from the "long" CH and C parameters are utilized in this study. Using the "generalized" or "short" scoring function values would result in a different MCMC sampling of ϵ and σ.

The optimal "long" CH parameter is on the boundary of the grid tested by Mick et al. Therefore, we do not have any scoring function values from simulation for $\epsilon_{\text{CH}} < 14$ K. A similar problem is faced for $\epsilon_{\text{C}} < 0.8$ K and $\sigma_{\text{C}} > 0.63$ nm but, fortunately, these regions are rarely sampled by the MCMC algorithm. To overcome the challenge of extrapolating outside of the domain where $S(\theta)$ is available, we fit $\ln(S(\theta))$ to a multi-variate normal distribution. This approach works well for the CH parameters because the CH scoring function has a fairly normal shape. While the assumption of normality is worse for S over the ϵ_{C}-σ_{C} parameter space, this does not significantly affect our results because of the infrequent sampling of this extrapolation region.

SI.III. MCMC validation

This section validates the combined bootstrap re-sampling and MCMC approach. Specifically, we compare the uncertainties (depicted as histograms) obtained in two different manners. First, where a single replicate simulation is performed for each MCMC-nb parameter set, which are pooled together for bootstrap re-sampling ($N_{\text{reps}} = 1$, $N_{\text{MCMC}} = 40$). Second, where 40 replicate simulations are performed for each set of MCMC-nb parameters, and bootstrap re-sampling is performed independently for each set of 40 replicates ($N_{\text{reps}} = 40$, $N_{\text{MCMC}} = 30$). Due to the large amount of simulations required for this comparison, we perform this analysis on a simpler system, namely, ethane at 137 K and saturated liquid density using the Potoff force field.

Figure SI.1 demonstrates that the uncertainties are nearly indistinguishable for the two methods. This provides empirical evidence that performing a single simulation with each MCMC parameter set is the same as performing numerous simulations with each MCMC parameter set. Also of interest is that the numerical uncertainties ($N_{\text{reps}} = 40$, $N_{\text{MCMC}} = 1$) are much smaller than the overall uncertainties, suggesting that parameter uncertainties play a large role for ethane.

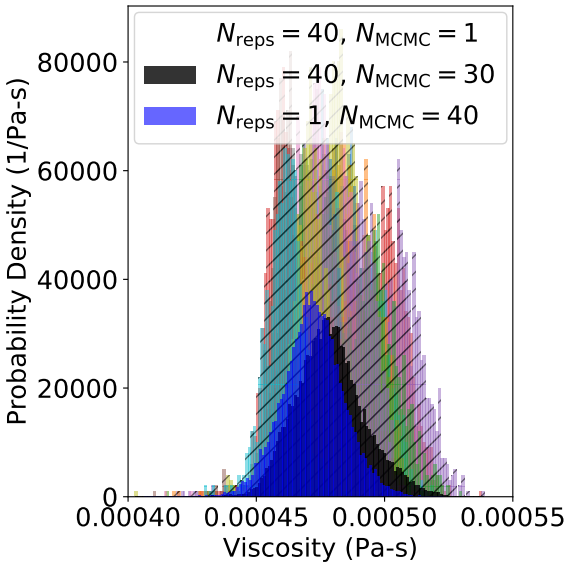

Figure SI.1: Validation of combined bootstrap re-sampling and MCMC approach utilized in study. Note that the uncertainties are almost indistinguishable between $N_{\text{reps}} = 1$, $N_{\text{MCMC}} = 40$ and $N_{\text{reps}} = 40$, $N_{\text{MCMC}} = 30$.

SI.IV. Torsion parameter uncertainty

In this section, we develop the skewed distribution for A_s, where the respective lower and upper 95 % confidence intervals correspond to -15 % and +40 % of the maximum torsional barrier. The viscosity values obtained with Potoff are considerably higher than those obtained with AUA4. Therefore, it is feasible, especially at higher pressures, that the optimal value of A_s is negative, i.e., the viscosity may be too high and, thus, decreasing the torsional barriers might improve the viscosity estimates. For this reason, unlike Nieto-Draghi et al., we consider $A_s < 0$.

To determine the appropriate scaling of the torsional barriers, Figure SI.2 presents a sensitivity analysis of η with respect to A_s. A_s is expressed as a percentage of the maximum for the non-shifted torsional potential. The viscosities in Figure SI.2 for 2,2,4-trimethylhexane are computed at 293 K and atmospheric pressure with 200 molecules and the Potoff force field. Also depicted is the only available experimental viscosity value at this temperature and pressure.

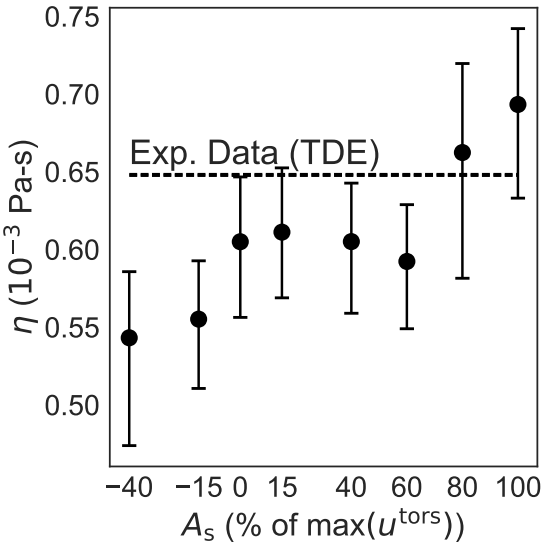

Figure SI.2: Sensitivity analysis of viscosity to torsional barrier heights. Simulations are performed at 293 K and atmospheric pressure. Experimental data are depicted as a dashed line. Uncertainties are expressed at 95 % confidence level, where the experimental uncertainty is approximately the line-width.

Figure SI.2 demonstrates that quantitative agreement with the experimental viscosity point necessitates an A_s value that is 80 % the maximum torsional barrier. Fearing some unforeseen consequences, we do not feel that obtaining quantitative agreement with this single experimental

value merits such a dramatic increase in the torsional barriers. For this reason, we adopt the largest percent increase proposed by Nieto-Draghi et al., i.e., 40 %.

By contrast, decreasing the torsional barriers by 40 % does have a significant impact on the predicted viscosity. We attribute this to the *gauche* barrier heights being approximately 40 % the *cis* barrier heights for the CH_i-CH_2-CH-CH_j torsional potential. Therefore, reducing all barriers by 40 % of the maximum torsional barrier nearly eliminates the equilibrium *gauche* conformations (see Figure 1 in the main text). Even a 15 % reduction has an appreciable effect on η. For this reason, we do not recommend reducing the barrier heights by more than 15 %.

SI.V. Green-Kubo integrals

Figure SI.3 presents the average Green-Kubo integral for all thirteen state points. Note that much longer simulations are required for high pressures/viscosities (bottom panel) than for low pressure/viscosities (top panel).

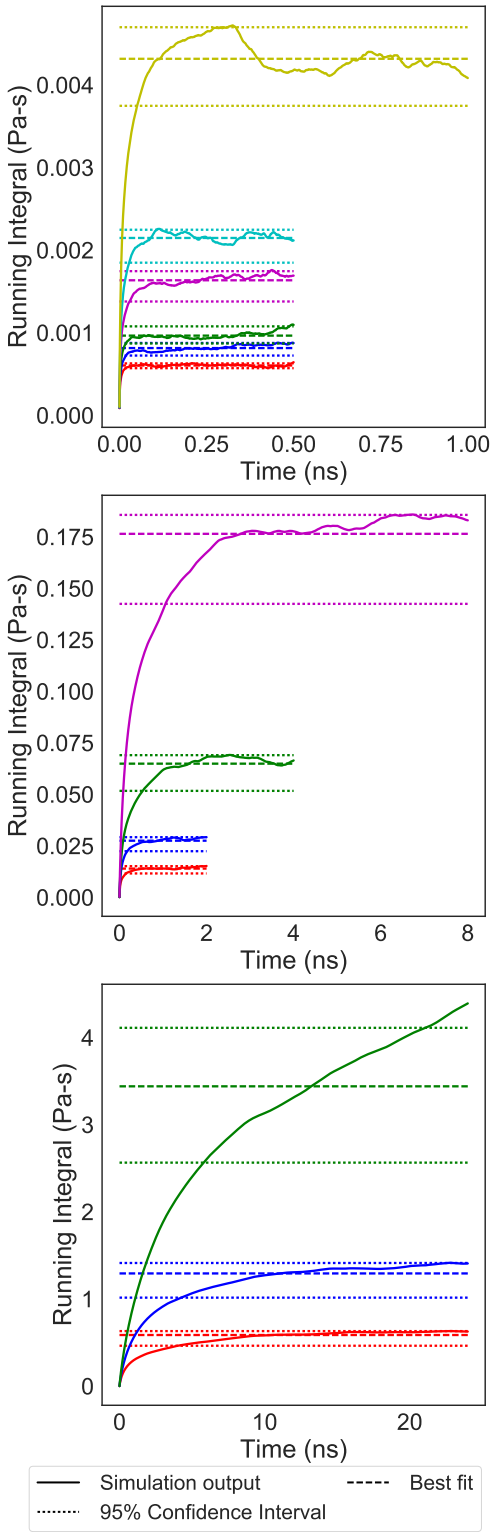

Figure SI.3: Green-Kubo integrals with respect to time. Top, middle, and bottom panels depict, respectively, low, intermediate, and high pressure/viscosity simulations where respective simulation times of 1 to 4 ns, 8 to 16 ns, and 24 to 48 ns are required to observe a Green-Kubo plateau.