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

SI.II.1. Theory

Mick et al. optimized the Potoff CH and C parameters using a scoring function (S) that weights the deviations for several different properties and their derivatives. MCMC requires an expression for the likelihood function (L) and, in particular, the log of the likelihood function ($\log_{10}(L)$). This section describes how we translate the scoring function into a log-likelihood function.

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

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

where D are the data, θ are the model parameters, n is the number of data points, σ is the standard deviation, and $\sum(y(\theta) - D_i)^2$ is the sum-squared-error (SSE). The log-likelihood can then be expressed as

$$\log_{10}(L(D|\theta)) \propto -\sum(y(\theta) - D_i)^2 \propto -SSE(\theta) \quad (2)$$

Clearly, minimizing the sum-squared-error is mathematically equivalent to maximizing the likelihood or log-likelihood when assuming the errors follow a normal distribution.

In MCMC, assuming a uniform prior, the probability of accepting a proposed or "new" parameter set (θ_{new}) given a previous or "old" parameter set (θ_{old}) is

$$\alpha = \frac{L(D|\theta_{\text{new}})}{L(D|\theta_{\text{old}})} \quad (3)$$

where α is the acceptance probability. For computational reasons, it is common to perform MCMC using the log-likelihood such that

$$\log_{10}(\alpha) = \log_{10}(L(D|\theta_{\text{new}})) - \log_{10}(L(D|\theta_{\text{old}})) \quad (4)$$

Note that all terms in Equation 1 that do not depend on θ cancel when computing $\log_{10}(\alpha)$. Assuming a normal distribution or, equivalently, minimizing SSE is achieved in MCMC by substi-

tuting Equation 2 into Equation 4 which yields

$$\log_{10}(\alpha) \quad = \quad \sum(y(\theta_{\text{old}}) \; - \; D_i)^2 \; - \; \sum(y(\theta_{\text{new}}) \; - \; D_i)^2 \quad = \quad SSE(\theta_{\text{old}}) \; - \; SSE(\theta_{\text{new}}) \quad (5)$$

note that the order of "new" and "old" changes due to the negative sign in Equation 2.

Because Potoff's scoring function (S) is not simply the sum-squared-error, minimizing S is not equivalent to maximizing the likelihood of a normal distribution. However, we can still apply the maximum likelihood criterion by substituting S for SSE in Equation 2

$$\log_{10}(L(D|\theta)) \propto -S(\theta) \quad (6)$$

and, therefore, the MCMC acceptance probability is simply

$$\log_{10}(\alpha) = S(\theta_{\text{old}}) - S(\theta_{\text{new}}) \quad (7)$$

SI.II.2. Implementation

With Equation 7, all that remains to perform MCMC is a way to compute S for any θ. This is achieved by interpolating the raw scoring function values 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 did not have any scoring function values 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 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 $\log_{10}(S(\theta))$ 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 of C, 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 saturation conditions.

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 parameter set is the same as performing numerous simulations with each 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 larger role for ethane.

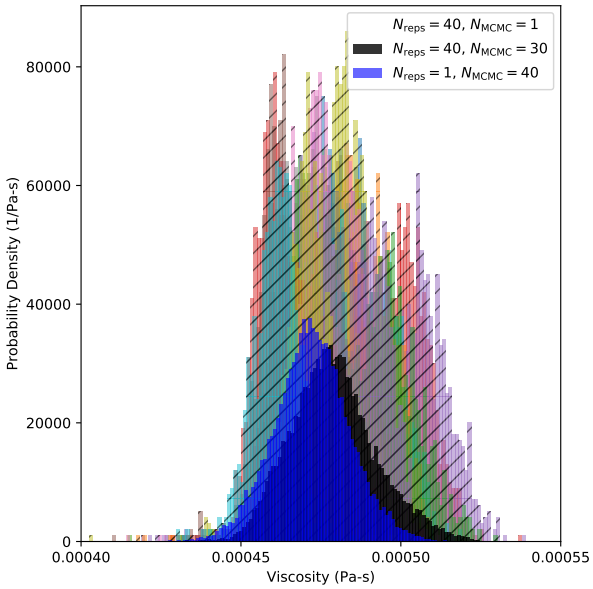

Figure SI.1: Validation of combined bootstrap re-sampling and MCMC approach utilized in study. Note that the uncertainties are almost indistinguishable between $N_{\text{reps}} = 1$, $N_{\text{MCMC}} = 40$ and $N_{\text{reps}} = 40$, $N_{\text{MCMC}} = 30$.

SI.IV. A_s distribution

In this section, we develop the skewed distribution for A_s, where the 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, 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. The viscosities in Figure SI.2 are computed at 293 K and atmospheric pressure. Also depicted is the only available experimental viscosity value.

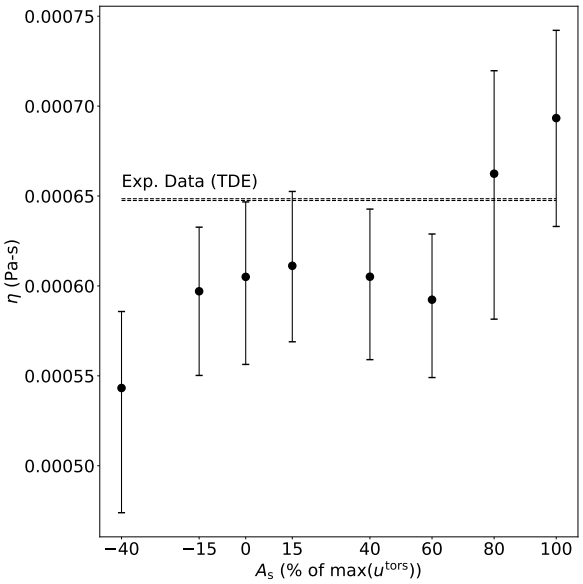

Figure SI.2: Sensitivity analysis of viscosity to torsional barrier heights. Simulations are performed at 293 K and atmospheric pressure. Experimental data are depicted as a dashed line. Uncertainties are expressed at 95 % confidence level. Experimental uncertainties are approximately the width of the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

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. Therefore, reducing all barriers by 40 % of the maximum torsional barrier nearly eliminates the equilibrium *gauche* conformations. For this reason, we adopt the smaller percent deviation proposed by Nieto-Draghi et al., i.e, 15 %, when reducing the barrier heights.

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 panel depicts lower pressure/viscosity simulations where 1 to 4 ns simulations are sufficiently long. Bottom panel depicts higher pressure/viscosity simulations where 8 to 48 ns simulations are required to observe a Green-Kubo plateau.

References

References