Bayesian inference analysis of transferable, united-atom, Mie λ-6 force fields for normal and branched alkanes. To be submitted to the Journal of Physical Chemistry, B.

Richard A. Messerly,[*,†] Michael R. Shirts,[*,‡] and Andrei F. Kazakov[*,†]

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

‡*Department of Chemical and Biological Engineering, University of Colorado, Boulder, Colorado, 80309*

E-mail: richard.messerly@nist.gov; michael.shirts@colorado.edu; andrei.kazakov@nist.gov

Abstract

Purpose

The aim of this study is to demonstrate, using Bayesian inference, that a UA Mie force field cannot adequately predict VLE and PVT of compressed liquids and supercritical fluids for normal and branched alkanes. For adequate prediction of VLE and compressed liquid pressures, we recommend using AUA or AA models.

1 Introduction

An accurate understanding of the relationship between pressure, volume (or density, ρ), and temperature (PVT) and caloric properties (such as heat capacity) for a given compound is essential for designing industrial chemical processes. Fundamental equations of state (FEOS), such as those based on the Helmholtz free energy, are a powerful approach for estimating PVT behavior and caloric properties. For example, the National Institute of Standards and Technology (NIST) REFPROP (Reference Fluid Properties) currently provides FEOS for around one hundred chemical species.[1] Unfortunately, most compounds do not have sufficient (reliable) experimental data covering a wide range of pressures, densities, and temperatures to develop a highly-accurate FEOS. Using an FEOS to extrapolate to temperatures and pressures that are significantly higher than those used in parameterizing the FEOS can result in large errors. Therefore, improvement in an FEOS at high temperatures and pressures necessitates additional data near those conditions.

The lack of experimental data at high temperatures and pressures, especially, is likely attributed to the inherent safety, cost, and complexity of such experiments. By contrast, molecular simulation (i.e. Monte Carlo, MC, and molecular dynamics, MD) methods at high temperatures and pressures do not suffer from any of these limitations. Therefore, in principle, molecular simulation could aid in developing FEOS.[2–7] For example, several

recent studies by Thol et al. supplement experimental data with molecular simulation results at temperatures and pressures beyond the range of available experimental temperatures and pressures.[8–10] Specifically, experimental data were available for temperatures and pressures up to 580 K and 130 MPa, 590 K and 180 MPa, and 560 K and 100 MPa for hexamethyldisiloxane, octamethylcyclotetrasiloxane, and 1,2-dichloroethane, respectively. Molecular simulations were performed for these compounds at temperatures and pressures up to 1200 K and 600 MPa, 1200 K and 520 MPa, 1000 K and 1200 MPa, respectively. The inclusion of these simulation results improved the performance of the FEOS at extreme temperatures and pressures.

Hydrocarbons are a fundamental feed-stock for many petrochemical processes and, therefore, large amounts of experimental data exist covering a wide range of PVT phase space. For these reasons, REFPROP contains highly-accurate FEOS for several hydrocarbons, most of which are shorter-chains (less than 20 carbons) with limited branching (i.e. only methyl branches). An appealing approach to develop FEOS for other hydrocarbons, is to utilize hybrid data sets consisting of experimental data and molecular simulation results at extreme temperatures and pressures.

The primary limitation for implementing molecular simulation at extreme temperatures and pressures is whether or not the force field, which is typically parameterized using VLE data, is reliable at those conditions, i.e. if the VLE optimal parameters are transferable to higher temperatures and pressures. In this study, we investigate how well the traditional force fields for predicting VLE extrapolate to higher temperatures (supercritical fluid) and pressures (compressed liquid). This analysis is performed for four normal and four branched alkanes by comparing the simulated compressibility factor (Z) with the REFPROP correlations, which are assumed to be reliable at these conditions.

The most accurate force fields for estimating hydrocarbon VLE properties (i.e. $\rho_{\mathrm{l}}^{\mathrm{sat}}$ and $P_{\mathrm{v}}^{\mathrm{sat}}$) are Transferable Potentials for Phase Equilibria (TraPPE)[11,12] (and, especially,

the recent TraPPE-2[13]), Errington,[14] anisotropic-united-atom (AUA4),[15,16] Potoff,[17,18] and Transferable anisotropic Mie potential (TAMie).[19,20] The TraPPE and Potoff force fields use a united-atom (UA) model while the TraPPE-2, Errington, AUA4, and TAMie force fields use an anisotropic-united-atom (AUA) model. Both a UA and AUA model group the hydrogen interactions with their neighboring carbon atom. However, the UA model assumes that the UA interaction site is that of the carbon atom, while an AUA model assumes that the AUA interaction site is shifted away from the carbon atom and towards the hydrogen atom(s). Although, in theory, an all-atom (AA) force field should yield more accurate results, from a parameterization standpoint, it is much easier to ensure that a global minimum is obtained when parameterizing UA and AUA force fields since fewer parameters are optimized simultaneously. The reduced computational cost is an additional benefit of the UA and AUA approach.

In addition to the classification of UA and AUA force fields, the existing force fields differ in the non-bonded functional form and corresponding parameters. The TraPPE, TraPPE-2, and AUA4 force fields use a Lennard-Jones (LJ) 12-6 potential, while the Potoff and TAMie force fields use the Mie λ-6 (or generalized Lennard-Jones) potential, and the Errington force field uses the Buckingham exponential-6 (Exp-6) potential. The three-parameter Mie λ-6 and Exp-6 potentials are more flexible than the two-parameter LJ 12-6 potential as the additional adjustable parameter controls the steepness of the repulsive barrier.

Previous work demonstrated that the UA LJ 12-6 potential cannot adequately estimate both $\rho_{\mathrm{l}}^{\mathrm{sat}}$ and $P_{\mathrm{v}}^{\mathrm{sat}}$ for *n*-alkanes.[21,22] For this reason, the TraPPE-UA force field was primarily developed to agree with saturated liquid densities.[11] By contrast, accurate prediction of both $\rho_{\mathrm{l}}^{\mathrm{sat}}$ and $P_{\mathrm{v}}^{\mathrm{sat}}$ over a wide temperature range is possible by varying the repulsive exponent of the LJ potential (i.e. the Mie λ-6 potential). Typically, the optimal value of λ is greater than 12 with a corresponding increase in the well depth (ϵ). Specifically for hy-

drocarbons, the Potoff UA force field uses $\lambda = 16$ while the TAMie force field uses $\lambda = 14$. However, there is some concern that increasing the repulsive exponent might have some undesirable consequences, especially at high pressures, where close range interactions will become more prevalent than at vapor-liquid equilibria. The purpose of this study is to determine whether or not the UA Mie potential is adequate for predicting both VLE and PVT at higher temperatures and pressures.

The outline for this manuscript is the following. Section 2 discusses the simulation and force field details. Section 3 is a case study for normal and branched alkanes using the existing force fields developed from VLE properties. Section 4 explains how Bayesian inference is employed to investigate the adequacy of the UA Mie potential. Section 5 presents the results from the Bayesian analysis. Section 7 reports the primary conclusions of this study.

2 Methods I

2.1 Simulation Details

We have selected four normal and four branched alkanes of varying chain-length and degree of branching. Specifically, we simulate ethane, propane, *n*-butane, *n*-octane, isobutane (2-methylpropane), isopentane (2-methylbutane), isohexane (2-methylpentane), isooctane (2,2,4-trimethylpentane), and neopentane (2,2-dimethylpropane).

Simulations for this study are performed in the NVT ensemble (constant number of molecules, N, constant volume, V, and constant temperature, T) using GROMACS version 2018.[23] Each simulation uses the Velocity Verlet integrator with a 2 fs time-step, 1.4 nm cut-off for non-bonded interactions with tail corrections for energy and pressure, Nosé-Hoover thermostat with a time constant of 1 ps, and fixed bond-lengths are con-

strained using LINCS with a LINCS-order of eight. The equilibration time was 0.1 ns for ethane and propane, 0.2 ns for *n*-butane, and 0.5 ns for all other compounds. The production time was 1 ns for ethane, 2 ns for propane and *n*-butane, and 4 ns for all other compounds. Replicate simulations were performed to validate that a single MD run of this length agrees with the average of several replicates, to within the combined uncertainty. A system size of 400 molecules is used for ethane, propane, and *n*-butane, while all other compounds use 800 molecules. Example input files are provided as Supporting Information.

Simulations are performed along a supercritical isotherm (with a reduced temperature, $T_\mathrm{r} \approx 1.2$) and five saturated liquid density isochores. Nine densities are simulated along the supercritical isotherm (T^IT) with five densities being those of the isochore densities. Two additional temperatures are simulated along each isochore, with one being the REFPROP saturation temperature (T^sat) and the inverse of the second isochore temperature is the average of $1/T^\mathrm{IT}$ and $1/T^\mathrm{sat}$. Thus, a total of 19 simulations are performed for each compound and force field. The specific state points for each compound studied are depicted in Figure 1, with the REFPROP saturation curve included as a reference. Tabulated values for the state points of each compound are provided in Supporting Information.

We use isothermal isochoric integration (ITIC) to convert the departure internal energies (U^dep) and compressibility factors (Z) obtained at the 19 state points to saturated VLE properties, namely, $\rho_\mathrm{l}^\mathrm{sat}$ and $P_\mathrm{v}^\mathrm{sat}$.[24,25] The equations for ITIC are:

$$\frac{A^\mathrm{dep}}{R_\mathrm{g}T^\mathrm{sat}} = \int_0^{\rho_\mathrm{l}^\mathrm{sat}} \frac{Z-1}{\rho}\partial\rho|_{T=T^\mathrm{IT}} + \int_{T^\mathrm{IT}}^{T^\mathrm{sat}} U^\mathrm{dep}\partial\left(\frac{1}{R_\mathrm{g}T}\right)|_{\rho=\rho_\mathrm{l}^\mathrm{sat}} \tag{1}$$

$$\rho_\mathrm{v}^\mathrm{sat} \approx \rho_\mathrm{l}^\mathrm{sat}\exp\left(\frac{A^\mathrm{dep}}{R_\mathrm{g}T^\mathrm{sat}} + Z_\mathrm{l}^\mathrm{sat} - 1 - 2B_2\rho_\mathrm{v}^\mathrm{sat} - 1.5B_3{\rho_\mathrm{v}^\mathrm{sat}}^2\right) \tag{2}$$

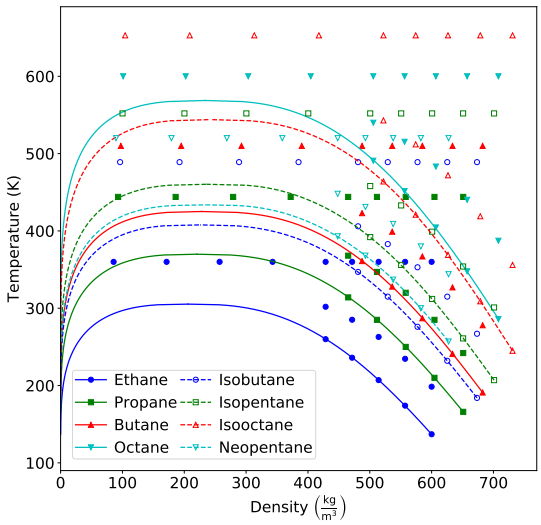

Figure 1: State points simulated for each compound studied. A total of 19 simulations are performed: nine densities along the supercritical isotherm and two temperatures along liquid density isochores. Filled symbols and solid lines correspond to *n*-alkanes, while empty symbols and dashed lines correspond to branched alkanes. The REFPROP saturation curve for each compound is included as a reference.

$$P_{\mathrm{v}}^{\mathrm{sat}} \approx (1 + B_2\rho_{\mathrm{v}}^{\mathrm{sat}} + B_3\rho_{\mathrm{v}}^{\mathrm{sat}^2})\rho_{\mathrm{v}}^{\mathrm{sat}} R_{\mathrm{g}} T^{\mathrm{sat}} \tag{3}$$

$$Z_{\mathrm{l}}^{\mathrm{sat}} = \frac{P_{\mathrm{v}}^{\mathrm{sat}}}{\rho_{\mathrm{l}}^{\mathrm{sat}} R_{\mathrm{g}} T^{\mathrm{sat}}} \tag{4}$$

where $A^{\mathrm{dep}} \equiv A - A^{\mathrm{ig}}$ is the Helmholtz free energy departure from ideal gas for temperature (T) equal to the saturation temperature (T^{sat}) and density (ρ) equal to the saturated liquid density $(\rho_{\mathrm{l}}^{\mathrm{sat}})$, $U^{\mathrm{dep}} \equiv U - U^{\mathrm{ig}}$ is the internal energy departure, $Z_{\mathrm{l}}^{\mathrm{sat}}$ is the saturated liquid compressibility factor (Z), B_2 is the second virial coefficient, B_3 is the third virial coefficient, T^{IT} is the isothermal temperature, and R_{g} is the universal gas constant. As discussed and validated in our previous work,[25] the B_2 and B_3 values found in Equations 2-3 are calculated using REFPROP correlations.[1] Details for this methodology are found in our previous work.[25]

2.2 Force field

A united-atom (UA) or anisotropic-united-atom (AUA) representation is used for each compound studied. The UA and AUA groups required for normal and branched alkanes are sp^3 hybridized CH_3, CH_2, CH, and C sites. For most literature models, a single (transferable) parameter set is assigned for each interaction site. However, two exceptions exist for the force fields studied. First, TAMie implements a different set of CH_3 parameters for ethane and other alkanes. Second, Potoff reports a "generalized" and "short/long" CH and C parameter set. The Potoff "generalized" parameter set is an attempt at a completely transferable set. However, since the "generalized" parameters performed poorly for some compounds, the "short/long" parameter set was proposed, where the "short" and "long" parameters are implemented when the number of carbons in the backbone is ≤ 4 and > 4, respectively.

A fixed bond-length is used for each bond between UA or AUA sites. Although TAMie is an AUA force field, only the terminal CH_3 sites have a displacement in the interaction site. This convention is much simpler to implement than other AUA approaches (such as AUA4) where non-terminal (i.e. CH_2 and CH) interaction sites also have a displacement distance. For this reason, we do not attempt to simulate the AUA4 force field for any compounds containing CH_2 and CH interaction sites. Therefore, the anisotropic shift in a terminal interaction site (i.e. CH_3) is treated simply as a longer effective bond-length (see Table 1). The bond-length for all non-terminal sites is 0.154 nm, except for the Errington Exp-6 force field which uses 0.1535 nm for CH_2-CH_2 bonds.

The angle and dihedral energies are computed using the same functional forms and parameters for each force field. Angular bending interactions are evaluated using a harmonic potential:

$$u^{\text{bend}} = \frac{k_\theta}{2}\left(\theta - \theta_0\right)^2$$

Table 1: Effective bond-lengths (nm) for terminal (CH_3) UA or AUA interaction sites. "Not-applicable" ("N/A") signifies that the force field either does not include these site types (e.g. Exp-6 and TraPPE-2) or that a more complicated notation than a simple effective bond-length is required to adequately represent the force field (i.e. AUA4).

Bond	TraPPE, Potoff	TAMie	Exp-6	AUA4	TraPPE-2
CH_3-CH_3	0.154	0.194	0.1839	0.1967	0.230
CH_3-CH_2	0.154	0.174	0.1687	N/A	N/A
CH_3-CH	0.154	0.174	N/A	N/A	N/A
CH_3-C	0.154	0.174	N/A	0.1751	N/A

where θ is the instantaneous bond angle, θ_0 is the equilibrium bond angle, and k_θ is the harmonic force constant which is equal to 62500 K/rad^2 for all potentials. Dihedral torsional interactions are determined using a cosine series:

$$u^{\text{tors}} = c_1[1 + \cos\phi] + c_2[1 - \cos 2\phi] + c_3[1 + \cos 3\phi]$$

where ϕ is the dihedral angle and c_i are the Fourier constants. The equilibrium bond angles and torsional parameters are found in Tables 2-3, respectively.

Table 2: Equilibrium bond angles (θ_0). x and y are values between 0-3.

Bending sites	θ_0 (degrees)
CH_x-CH_2-CH_y	114.0
CH_x-CH-CH_y	112.0
CH_x-C-CH_y	109.5

Non-bonded interactions between two different molecules and united-atom sites separated by more than three bonds are calculated using either a Lennard-Jones 12-6, Mie λ-6, or Buckingham Exponential-6 potential. The Mie λ-6 potential is:

$$u^{\text{vdw}}(\epsilon, \sigma, \lambda; r) = \left(\frac{\lambda}{\lambda - 6}\right)\left(\frac{\lambda}{6}\right)^{\frac{6}{\lambda-6}} \epsilon \left[\left(\frac{\sigma}{r}\right)^{\lambda} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{5}$$

Table 3: Fourier constants (c_i) in K. x and y are values between 0-3.

Torsion sites	c_0	c_1	c_2	c_3
CH_x-CH_2-CH_2-CH_y	0.0	355.03	-68.19	791.32
CH_x-CH_2-CH-CH_y	-251.06	428.73	-111.85	441.27
CH_x-CH_2-C-CH_y	0.0	0.0	0.0	461.29
CH_x-CH-CH-CH_y	-251.06	428.73	-111.85	441.27

where u^{vdw} is the van der Waals interaction, σ is the distance (r) where $u^{\text{vdw}} = 0$, $-\epsilon$ is the energy of the potential at the minimum $\left(\text{i.e. } u^{\text{vdw}} = -\epsilon \text{ and } \frac{\partial u^{\text{vdw}}}{\partial r} = 0 \text{ for } r = r_{\min}\right)$, and λ is the repulsive exponent.

Note that the Mie λ-6 potential reduces to the LJ 12-6 potential for $\lambda = 12$. Therefore, the LJ 12-6 potential can be considered a special subclass of the Mie λ-6 potential. It is important to mention that, although an attractive exponent of 6 has a strong theoretical basis, $\lambda = 12$ is a historical artifact that was chosen primarily for computational purposes.[26] For the same reason (i.e. computational efficiency), a common practice to date is to use integer values of λ in Equation 5. The non-bonded force field parameters for TraPPE (and TraPPE-2), Potoff, AUA4, and TAMie are provided in Table 4.

Non-bonded interactions between two different site types (i.e. cross-interactions) are determined using Lorentz-Berthelot combining rules[26] for ϵ and σ with an arithmetic mean for the repulsive exponent (λ) (as recommended by Potoff and Bernard-Brunel[17]):

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}} \tag{6}$$

$$\sigma_{ij} = \frac{\sigma_{ii} + \sigma_{jj}}{2} \tag{7}$$

$$\lambda_{ij} = \frac{\lambda_{ii} + \lambda_{jj}}{2} \tag{8}$$

where the ij subscript refers to cross-interactions and the subscripts ii and jj refer to

Table 4: Non-bonded (intermolecular) parameters for TraPPE[11,12] (and TraPPE-2[13]), Potoff,[17,18] AUA4,[15,27] and TAMie[19,20] force fields. The "short/long" Potoff CH and C parameters are included in parenthesis. The ethane specific parameters for TAMie are included in parenthesis.

	TraPPE (TraPPE-2)			Potoff (S/L)		
United-atom	ϵ (K)	σ (nm)	λ	ϵ (K)	σ (nm)	λ
CH_3	98 (134.5)	0.375 (0.352)	12	121.25	0.3783	16
CH_2	46	0.395	12	61	0.399	16
CH	10	0.468	12	15 (15/14)	0.46 (0.47/0.47)	16
C	0.5	0.640	12	1.2 (1.45/1.2)	0.61 (0.61/0.62)	16
	AUA4			TAMie		
CH_3	120.15	0.3607	12	136.318 (130.780)	0.36034 (0.36463)	14
CH_2	86.29	0.3461	12	52.9133	0.40400	14
CH	50.98	0.3363	12	14.5392	0.43656	14
C	15.04	0.244	12	N/A	N/A	N/A

same-site interactions.

3 Case study for alkanes

The purpose of this case study is to demonstrate that the existing UA and AUA force fields for normal and branched alkanes that were parameterized with VLE properties do not predict the proper PVT behavior at higher temperatures and pressures (with the exception of ethane for the TraPPE-2 potential). Figures 2 and 3 plot the compressibility factor with respect to inverse temperature for *n*-alkanes and branched alkanes, respectively. Note that saturation corresponds to $Z \approx 0$ for each isochore.

The "Potoff" results in Figure 3 are only for the the "short/long" model, since the "short/long" model is more accurate than the "generalized" model. The results for the "generalized" model do not provide any additional insight but are found in the Supporting Information.

Figure 2 demonstrates that the existing literature force fields for *n*-alkanes, while accurate for VLE, do not capture the correct PVT behavior at high pressures, i.e. the higher temperatures and highest isochore densities (ρ_0 and ρ_1). Figure 3 shows that these force fields are typically less reliable at VLE for branched alkanes than for *n*-alkanes but, more importantly, the same erroneous trend in Z is observed as in Figure 2.

The one exception is the TraPPE-2 model for ethane, which reproduces the entire PVT phase space simulated. This result is a bit surprising considering the TraPPE-2 model has only three fitting parameters (ϵ, σ, and the effective bond-length) while the TAMie model has an additional fitting parameter (λ). It is important to note that TraPPE-2 uses a much longer effective bond-length of 0.230 nm while TAMie did not consider bond-lengths larger than 0.194 nm (see Table 1). Therefore, the fact that the TraPPE-2 force field extrapolates to high pressures better than TAMie suggests that, at high pressures, it is important to account for hydrogens either explicitly (AA model) or with a longer effective bond-length than that typically used for AUA models. It is also possible that a four parameter optimization, such as that used by TAMie, is over fit to the VLE data and would perform better if high pressure PVT data were included in the parameterization.

In general, a clear bias is observed for the LJ 12-6 potentials (TraPPE-UA and AUA4) and the Mie λ-6 potentials (Potoff and TAMie). Specifically, the LJ 12-6 and Mie λ-6 potentials under and over predict Z at high pressures, respectively. These results make intuitive sense as the repulsive barriers are steeper for the respective Mie 16-6 and 14-6 potentials of the Potoff and TAMie force fields. Another surprising trend is that the Errington (AUA Exp-6) model also has a positive bias at high pressures, suggesting that an exponential repulsive barrier is also too steep. Unfortunately, a direct comparison of the non-bonded potentials for AUA models is difficult because each model has a different anisotropic displacement. By contrast, a comparison of TraPPE-UA and Potoff is straightforward because they use the same bond-lengths and the same non-bonded potential (Equation

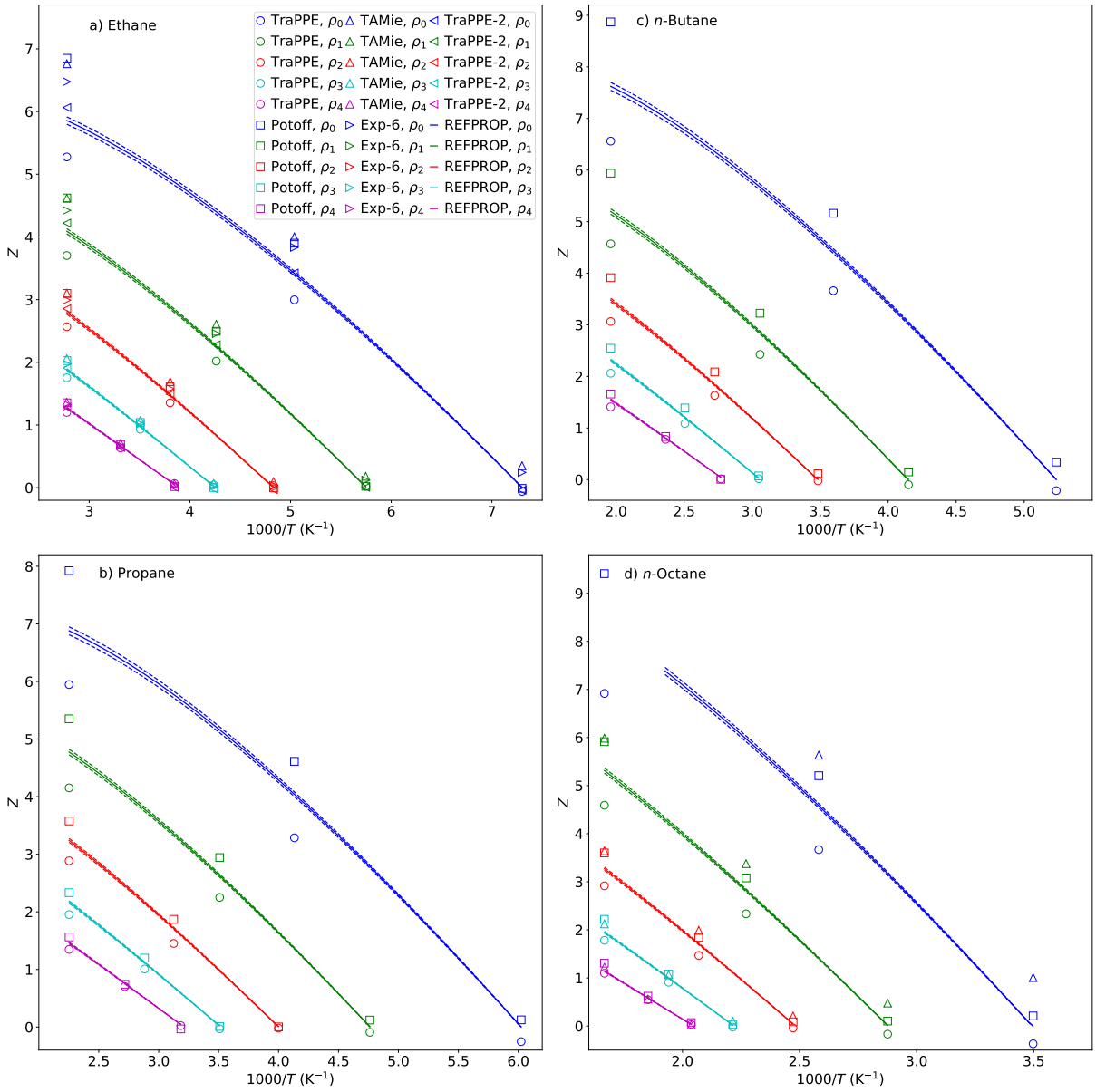

Figure 2: Compressibility factors (Z) along isochores agree at saturation $(Z \approx 0)$ but deviate strongly at higher pressures. Densities are distinguished by color and are labeled such that $\rho_0 > \rho_1 > \rho_2 > \rho_3 > \rho_4$. Panels a)-d) correspond to ethane, propane, n-butane, and n-octane, respectively. TraPPE and Potoff simulation results are depicted using open circles and squares, respectively, with error bars representing two times the standard deviation of the fluctuations from a single simulation. Solid lines represent REFPROP correlations, with dashed lines representing a 1% uncertainty in REFPROP values. Simulation error bars are approximately one symbol size.

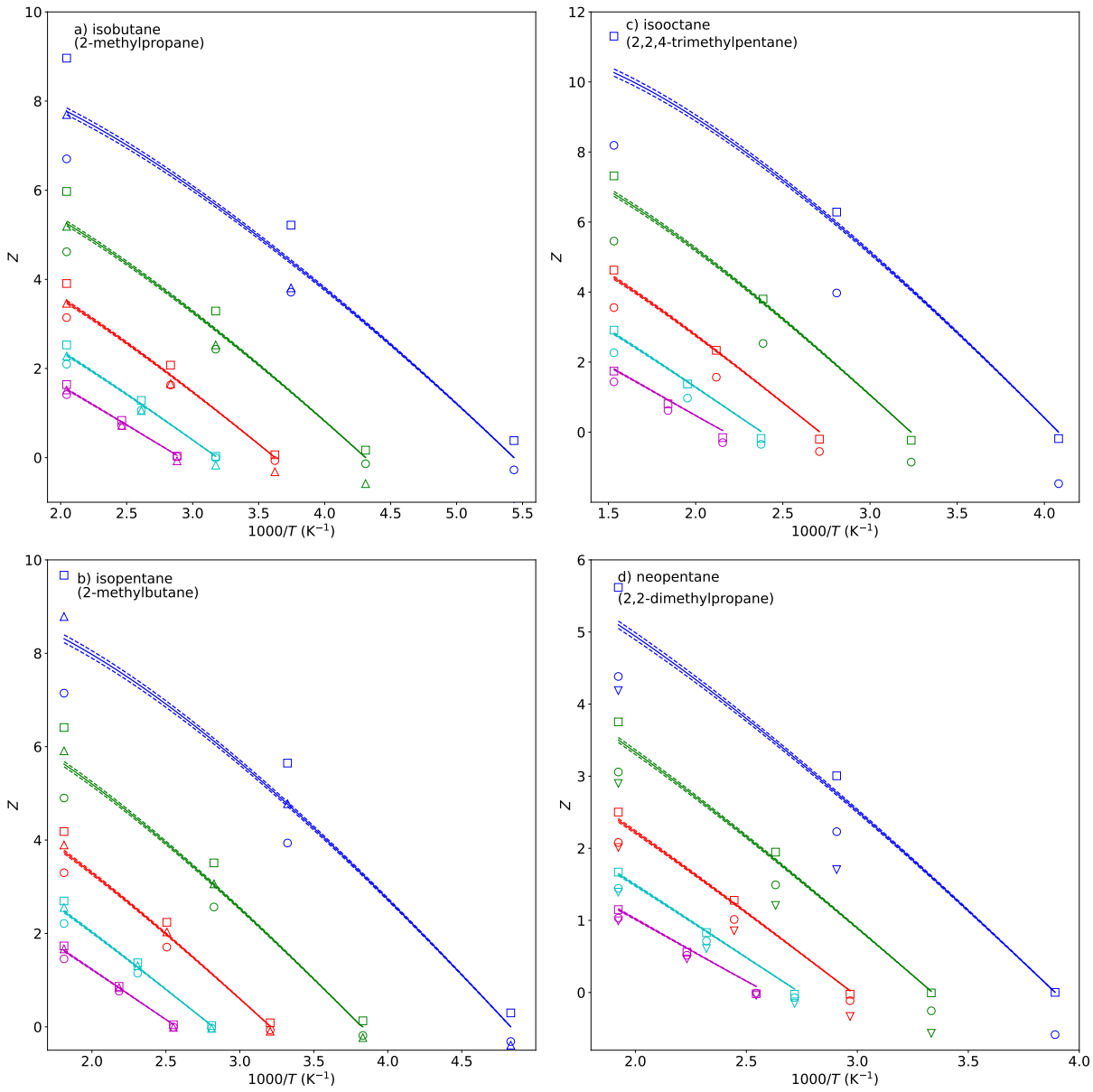

Figure 3: Compressibility factors (Z) along isochores for branched alkanes are not as accurate as normal alkanes at saturation $(Z \approx 0)$ and deviate strongly at higher pressures. Panels a)-d) correspond to isobutane, isopentane, isooctane, and neopentane, respectively. Symbols, lines, uncertainties, and formatting are the same as those in Figure 2.

5). For this reason, the remainder of this document focuses on the united-atom Mie λ-6 potentials where all bond-lengths are 0.154 nm.

Since the TraPPE-UA (LJ 12-6) potential under predicts Z and the Potoff (UA Mie 16-6) potential over predicts Z, it seems reasonable that a UA Mie 13-6, 14-6, or 15-6 model would demonstrate the proper trend, if parameterized appropriately. However, as demonstrated in Section 5, there does not exist a set of ϵ, σ, and λ that reasonably predicts $\rho_{\mathrm{l}}^{\mathrm{sat}}$, $P_{\mathrm{v}}^{\mathrm{sat}}$, and PVT of supercritical fluids and compressed liquids for the UA model. To understand this point, it is important to remember that the UA LJ 12-6 (TraPPE-UA) force field cannot adequately predict both $\rho_{\mathrm{l}}^{\mathrm{sat}}$ and $P_{\mathrm{v}}^{\mathrm{sat}}$. In other words, determining the optimal value of λ for predicting PVT of supercritical fluids and compressed liquids does not guarantee accurate prediction of $P_{\mathrm{v}}^{\mathrm{sat}}$. See Section 5 for a further discussion.

4 Methods II

The results presented in Section 3 demonstrate that none of the existing force fields studied reproduce the PVT behavior for supercritical fluids and compressed liquids. However, recall that each of these force fields was parameterized using only VLE properties. Therefore, it is possible that including both VLE and PVT properties in the parameterization objective function will improve the results. However, if no combination of ϵ, σ, and λ is capable of predicting VLE properties and PVT behavior, we can conclude that the UA Mie λ-6 potential is inadequate for this purpose and, therefore, should not be used when developing FEOS with molecular simulation results.

In order to rigorously quantify if the UA Mie λ-6 potential is "adequate", we perform a Bayesian inference analysis. We refer the reader to the literature for a thorough discussion of Bayesian statistics. In Section 4.1, we review some basic concepts of Bayes theorem, we define the posterior, likelihood, and prior functions, and we discuss the Markov Chain Monte Carlo (MCMC) sampling approach. As MCMC can be computationally burdensome, especially when coupled with molecular simulations, we use Multistate Bennett

Acceptance Ratio (MBAR) as a surrogate model to reduce the computational cost for determining the VLE properties $\rho_{\mathrm{l}}^{\mathrm{sat}}$ and $P_{\mathrm{v}}^{\mathrm{sat}}$ and Z of the supercritical fluids and compressed liquids (see Section 4.2).

The Bayesian inference analysis for CH_3 and CH_2 sites is performed sequentially. Specifically, rather than sampling from a four-dimensional (i.e. $\epsilon_{CH_3}, \epsilon_{CH_2}, \sigma_{CH_3}, \sigma_{CH_2}$ for a given value of λ_{CH_3} and λ_{CH_2}) Markov Chain, we implement a sequential two-dimensional approach by assuming the CH_3 parameters from ethane are transferable to propane, *n*-butane, and *n*-octane. As mentioned in Section 2.2, it is common to limit λ to integer values. Although sampling from discrete values is possible with MCMC, because of the strong correlation between ϵ and λ advanced sampling methods are required to achieve good acceptance ratios when varying λ by an integer amount. For this reason, we perform the MCMC analysis using fixed values of λ. This approach is computationally more efficient since we are only concerned with a few values of λ (i.e. 12-18). Finally, since the CH_3 and CH_2 results are similar and suggest that a UA Mie λ-6 potential cannot estimate both VLE and PVT, we did not find it necessary to repeat this process for the CH and C interaction sites.

4.1 Bayesian Analysis

Bayesian inference is used to quantify the uncertainty in the non-bonded parameters (ϵ and σ) and to determine the evidence for different values of λ. Bayes theorem states that

$$Pr(\theta|D) = \frac{Pr(D|\theta)Pr(\theta)}{Pr(D)} \tag{9}$$

where Pr denotes a probability distribution function, θ is the parameter set (i.e. ϵ and σ for a given Mie λ-6 potential), and D are the data. $Pr(\theta|D)$ is commonly referred to as the "posterior, $Pr(D|\theta)$ is the "likelihood" (alternatively expressed as $L(\theta|D)$), $Pr(\theta)$ is the

"prior", and $Pr(D)$ is a normalization constant. The evidence for different values of λ is determined by integrating the numerator of Equation 9 for all values of ϵ and σ. Note that this can be viewed as marginalizing the three-dimensional distribution with respect to λ. The Bayes factor for two different values of λ is obtained from the ratio of the respective evidences.

Markov Chain Monte Carlo (MCMC) is the traditional approach for numerically sampling from the probability distribution $Pr(\theta|D)$. A Markov Chain is created by proposing new ϵ or σ values and accepting those moves based on the ratio of the probability between the previous parameter set and the proposed parameter set:

$$\alpha = \min\left(1, \frac{Pr(\theta_{i+1}|D)}{Pr(\theta_i|D)}\right) \tag{10}$$

where α is the acceptance probability, θ_i is the previous parameter set, and θ_{i+1} is the proposed parameter set. The amount to which ϵ or σ is varied ($\delta\epsilon$ and $\delta\sigma$) for each MCMC step is tuned such that approximately $\frac{1}{3}$ of the moves are accepted. This "tuning" period (also referred to as a "burn-in" period) is followed by a production period where $\delta\epsilon$ and $\delta\sigma$ do not change. Details for MCMC are provided in Supporting Information (i.e. number of steps for burn-in and production, frequency that step sizes are updated, resulting acceptance percentages, etc.).

Because MCMC moves are accepted based on Equation 10 and the denominator in Equation 9 (i.e $(Pr(D))$ does not depend on θ, the acceptance probability is independent of $Pr(D)$. Also, we use a "non-informative prior" such that the acceptance probability is independent of $Pr(\theta)$ (although we do include a lower bound that the parameters are positive, i.e. $Pr(\theta)$ is uniform for all values of ϵ, σ, and λ greater than 0). Therefore, the probability of accepting θ_{i+1} is based completely on the likelihood. For this reason, we discuss in some detail how we calculate $L(\theta|D)$.

The likelihood is calculated using a multi-variable normal distribution. The variance accounts for the uncertainties of both the experimental data and the computational analysis (i.e. the methods discussed in Section 4.2). The uncertainties are assumed to be independent such that the combined variance is the sum of the experimental and computational variances.[28]

The parameter sets sampled from MCMC (θ_{MCMC}) provide an estimate of the uncertainty in θ (i.e. ϵ and σ). Figure 4 in Section 5 depicts the uncertainty in the parameters. This parameter uncertainty propagates when estimating another property (q), which may or may not be included in D. For example, although D only consists of VLE properties ($\rho_{\mathrm{l}}^{\mathrm{sat}}$ and $P_{\mathrm{v}}^{\mathrm{sat}}$, specifically) we also propagate the uncertainties in ϵ and σ to Z at high temperatures and pressures by implementing posterior prediction. The probability distribution of q ($Pr(q|D)$) is often approximated by developing a histogram of q for the MCMC parameter sets, i.e. $q(\theta_{\mathrm{MCMC}})$. Since a large number of MCMC samples are required for adequate representations of $Pr(\theta|D)$ and $Pr(q|D)$, MCMC is computationally infeasible when a direct molecular simulation is required for every θ_{MCMC}. For this reason, a surrogate model is used to approximate $L(\theta|D)$ (and, thereby, $Pr(\theta|D)$) and $q(\theta_{\mathrm{MCMC}})$ (and, thereby, $Pr(q|D)$).

4.2 Surrogate Model

A typical Markov Chain requires $O(10^4\text{-}10^5)$ Monte Carlo steps, where the likelihood function must be evaluated at each step. Since $L(\theta|D)$ depends on the force field parameters (ϵ, σ), an MCMC approach is computationally infeasible if computing $L(\theta|D)$ requires performing direct molecular simulations for every proposed set of ϵ and σ. For this reason, surrogate models are an essential tool for Bayesian methods such as MCMC. We use a configuration-sampling-based surrogate model, where configurations are sampled using a small group of reference parameter sets $(\epsilon_{\mathrm{ref}}, \sigma_{\mathrm{ref}})$. Ensemble averages for the

MCMC parameter sets (θ_{MCMC}) are estimated by reweighting the sampled reference configurations using Multistate Bennett Acceptance Ratio (MBAR). The properties that are estimated using MBAR are the departure internal energy (U^{dep}) and the compressibility factor (Z).

As discussed in Section 2, we use Isothermal Isochoric Integration (ITIC) to convert the MBAR estimated U^{dep} and Z values at the 19 ITIC state points to saturated liquid densities ($\rho_{\mathrm{l}}^{\mathrm{sat}}$) and vapor pressures ($P_{\mathrm{v}}^{\mathrm{sat}}$), since $\rho_{\mathrm{l}}^{\mathrm{sat}}$ and $P_{\mathrm{v}}^{\mathrm{sat}}$ are the data (D) included in $L(\theta|D)$. Details for the implementation of MBAR and ITIC (MBAR-ITIC) is discussed elsewhere.[25]

The ITIC analysis provides VLE properties at only 5 saturation temperature values (T_{ITIC}^{sat}), while the experimental data set (D) may have hundreds of saturation temperatures ($T_{\mathrm{exp}}^{\mathrm{sat}}$). Although we could use empirical correlations fit to the experimental data (i.e. REFPROP, ThermoData Engine (TDE)[29]), raw experimental data are preferred when performing a Bayesian analysis. For this reason, we instead use empirical model fits to interpolate the ITIC VLE properties ($\rho_{\mathrm{l,ITIC}}^{\mathrm{sat}}$) and ($P_{\mathrm{v,ITIC}}^{\mathrm{sat}}$) at any saturation temperature. Specifically, we fit $P_{\mathrm{v,ITIC}}^{\mathrm{sat}}$ to the Antoine equation:

$$\log_{10}(P_{\mathrm{v}}^{\mathrm{sat}}) = a_0 + \frac{a_1}{T^{\mathrm{sat}} + a_2} \tag{11}$$

where a_i are fitting parameters. We fit $\rho_{\mathrm{l,ITIC}}^{\mathrm{sat}}$ to a combined rectilinear and density scaling law expression:[22]

$$\rho_{\mathrm{l}}^{\mathrm{sat}} = b_0 + b_1(b_2 - T^{\mathrm{sat}}) + b_3(b_2 - T^{\mathrm{sat}})^{b_4} \tag{12}$$

where b_i are fitting parameters, although we use a fixed value of $b_4 = 0.326$ that is common for simple fluids. b_0 and b_2 are rough estimates of the critical density (ρ_c) and critical temperature (T_c). More reliable estimates of the critical point require a similar expression for $\rho_{\mathrm{v}}^{\mathrm{sat}}$, but this is unnecessary for our purposes since we are only including $\rho_{\mathrm{l}}^{\mathrm{sat}}$ in D.

In summary, MBAR, ITIC, and Equations 11-12 enable prediction of $\rho_{\mathrm{l}}^{\mathrm{sat}}$ and $P_{\mathrm{v}}^{\mathrm{sat}}$ for any ϵ and σ by performing direct simulations with only a few reference parameter sets. In addition, since the Mie potential (Equation 5) can be expressed as a linear equation with respect to r^{-6} and $r^{-\lambda}$, we implement basis functions to efficiently recompute the energies and forces that are required for MBAR and ITIC (for details see Appendix of Messerly et al.[25]). In total, this methodology reduces the computational cost for computing the likelihood based on VLE data by several orders of magnitude compared to direct simulation using Gibbs Ensemble Monte Carlo (GEMC) or Grand Canonical Monte Carlo (GCMC) histogram reweighting (HR).

Quantifying the uncertainty due to MBAR, ITIC, and Equations 11-12 (i.e. u_{c}) is essential for evaluating $L(\theta|D)$. Rather than performing a rigorous statistical assessment, we use an empirical approach for estimating u_{c}. Specifically, we compare our estimated $\rho_{\mathrm{l}}^{\mathrm{sat}}$ and $P_{\mathrm{v}}^{\mathrm{sat}}$ values for TraPPE and Potoff with those reported in the literature obtained using Gibbs Ensemble Monte Carlo (GEMC) or Grand Canonical Monte Carlo (GCMC) histogram reweighting (HR). This comparison has the added benefit that it incorporates possible deviations associated with the simulation package, and post-simulation analysis. For example,

The resulting error model estimates u_{c} to be around 1% and 5% for $\rho_{\mathrm{l}}^{\mathrm{sat}}$ and $P_{\mathrm{v}}^{\mathrm{sat}}$, respectively. For the compounds investigated in this study, these uncertainties are much larger than the experimental uncertainties (u_{exp}). Since the size of the parameter space sampled by MCMC depends almost entirely on u_{c}, we use a conservative estimate for u_{c}. In other words, the θ_{MCMC} sampled points are the only feasible values of ϵ and σ for optimizing $\rho_{\mathrm{l}}^{\mathrm{sat}}$ and $P_{\mathrm{v}}^{\mathrm{sat}}$.

5 Results

Figure 4 presents the MCMC sampled ϵ_{CH_2} and σ_{CH_2} parameter sets with Panels a) and b) corresponding to $\lambda_{CH_2} = 16$ and $\lambda_{CH_2} = 14$, respectively. Panel a) contains the MCMC parameter sets for propane, *n*-butane, and *n*-octane, while Panel b) contains results for propane and *n*-butane. Figure 4 also includes contours of the average percent deviations (AD%) in P^{high} relative to the REFPROP correlations, with the "REFPROP uncertainty" region corresponding to AD% of ± 1.

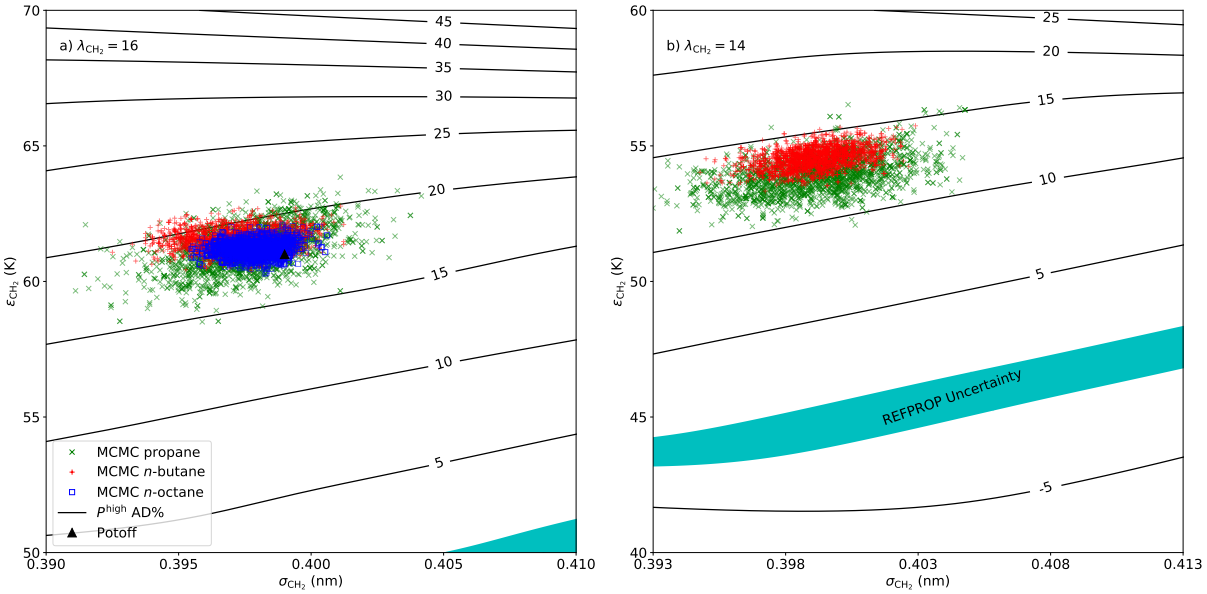

Figure 4: MCMC sampled ϵ_{CH_2} and σ_{CH_2} parameter sets result in large AD% for P^{high}. Panels a) and b) correspond to $\lambda_{CH_2} = 16$ and $\lambda_{CH_2} = 14$, respectively. REFPROP uncertainty in P^{high} is $\pm 1\%$. Potoff parameter set is provided as a reference for $\lambda_{CH_2} = 16$.

Notice in Figure 4 that the MCMC sampled ϵ_{CH_2} and σ_{CH_2} parameter sets, for a given value of λ_{CH_2}, overlap considerably for the different compounds. This supports the common assumption of transferability of CH_2 parameters between different *n*-alkanes. Also, note that the uncertainty in the parameters is largest for propane and smallest for *n*-octane. This suggests that, as expected, the sensitivity of ρ_l^{sat} and P_v^{sat} with respect to the CH_2 parameters increases with increasing number of CH_2 interaction sites. Notice, in

Panel a), that the Potoff parameter set for $\lambda_{\mathrm{CH_2}} = 16$ is within the MCMC sample region.

More importantly, for the purposes of this manuscript, the MCMC sampled $\epsilon_{\mathrm{CH_2}}$ and $\sigma_{\mathrm{CH_2}}$ parameter sets have large AD% in P^{high}. Specifically, $\lambda_{\mathrm{CH_2}} = 16$ and $\lambda_{\mathrm{CH_2}} = 14$ have, respectively, AD% of $\approx$ 16-21 and $\approx$ 10-15, much greater than the REFPROP uncertainty of around 1%. Because the "REFPROP uncertainty" contours are roughly parallel to the MCMC region and found at much lower $\epsilon_{\mathrm{CH_2}}$ (around 45 K for $\sigma_{\mathrm{CH_2}} = 0.399$ nm), in order to accurately predict P^{high}, it is necessary to sacrifice accuracy in ρ_{l}^{sat} and $P_{\mathrm{v}}^{\mathrm{sat}}$. This suggests that neither the UA Mie 16-6 or 14-6 models are capable of predicting VLE and PVT for supercritical fluids and compressed liquids of *n*-alkanes. Finally, although the UA Mie 14-6 AD% is slightly better than the UA Mie 16-6 AD%, recall that the UA Mie 14-6 is significantly less reliable for VLE. Therefore, considering the deprecation in VLE, the marginal gain in accuracy for P^{high} likely does not merit using a UA Mie 14-6 potential.

1. Figure: The uncertainty regions for CH3, CH2, CH, and C. I can include 14-6, 15-6, and 16-6. Perhaps I will only do this rigorous analysis for CH3 or for CH3 and CH2. Probably not for all. I could include the results from the alternative posterior (excluding Pvsat and including high pressures) but then it might be out of place in this section.

2. Bayes factors demonstrate that, for VLE, a 15-6 or 16-6 potential are favored significantly more than a 14-6 (could include 17-6 or 18-6 as well)

3. CH2 credible regions overlap considerably between propane, n-butane, and n-octane

4. By comparing the Bayes factor of a transferable CH2 site and three independent CH2 sites we observe that the CH2 sites are indistinguishable

5. Statement about CH credible regions for isobutane, isopentane, and isohexane

6. Statement about C credible region for isooctane and neopentane

6 Recommendations and Limitations

Note that the simulation values used by Thol et al. were derivatives of the residual Helmholtz free energy ($\partial^n a^r$) with respect to inverse temperature and/or density,[8–10] while in this study we simply compare the PVT behavior. Aside from the advantage of simplicity (most simulation packages do not provide $\partial^n a^r$), this choice is based on the fact that PVT is more readily understood and easier to visualize. In other words, it is easier to quantify the impact on process design caused by deviations in PVT behavior than derivatives in the residual Helmholtz free energy. Furthermore, as demonstrated by Thol et al., an inaccurate prediction of some $\partial^n a^r$ does not necessarily result in poor prediction of PVT behavior or heat capacities.[8] It is important to remember that PVT depends only on the first derivative of Helmholtz free energy with respect to density. Therefore, future work should investigate the adequacy of force fields to predict heat capacities, which depend on temperature derivatives, at higher temperatures and pressures. We would like to emphasize that, although we did not use $\partial^n a^r$ for our analysis, including higher order derivatives of the residual Helmholtz free energy from molecular simulation has significant advantages for developing FEOS as it eliminates redundant information found in traditional macroscopic properties.[2–7]

7 Conclusions

Recently, molecular simulation results at extreme temperatures and pressures have been used to supplement experimental data when developing a fundamental equation of state. As discussed by Thol et al., due to uncertainties and deficiencies in the force field, experimental data should be favored over molecular simulation values whenever possible. However, in principle, a FEOS could be developed for compounds without any

experimental data by using only molecular simulation results, if the force field were reliable and transferable over different PVT conditions. In part, one of our aims was to determine whether the united-atom Mie λ-6 potential for normal and branched alkanes was reliable enough that a FEOS could be developed strictly from molecular simulation results. Unfortunately, the Bayesian statistical analysis performed in this study suggests that this model type (UA Mie λ-6) is not adequate for predicting both VLE properties and high pressures for supercritical fluids and compressed liquids. Specifically, no set of ϵ, σ, and λ can adequately predict VLE and PVT behavior. Therefore, we recommend that alternative models be considered for developing FEOS, such as force fields using anisotropic-united-atom, all-atom, and/or alternative non-bonded potentials, e.g. Buckingham exponential-6, extended Lennard-Jones, etc.

References

(1) Lemmon, E. W.; Huber, M. L.; McLinden, M. O. NIST Standard Reference Database 23: Reference Fluid Thermodynamic and Transport Properties-REFPROP, Version 9.1, National Institute of Standards and Technology. 2013; https://www.nist.gov/srd/refprop.

(2) Thol, M.; Rutkai, G.; Köster, A.; Lustig, R.; Span, R.; Vrabec, J. *Journal of Physical and Chemical Reference Data* **2016**, *45*, 023101.

(3) Thol, M.; Rutkai, G.; Span, R.; Vrabec, J.; Lustig, R. *International Journal of Thermophysics* **2015**, *36*, 25–43.

(4) Rutkai, G.; Thol, M.; Span, R.; Vrabec, J. *Molecular Physics* **2017**, *115*, 1104–1121.

(5) Lustig, R.; Rutkai, G.; Vrabec, J. *Molecular Physics* **2015**, *113*, 910–931.

(6) Rutkai, G.; Thol, M.; Lustig, R.; Span, R.; Vrabec, J. *The Journal of Chemical Physics* **2013**, *139*, 041102.

(7) Rutkai, G.; Vrabec, J. *Journal of Chemical & Engineering Data* **2015**, *60*, 2895–2905.

(8) Thol, M.; Dubberke, F.; Rutkai, G.; Windmann, T.; Köster, A.; Span, R.; Vrabec, J. *Fluid Phase Equilibria* **2016**, *418*, 133 – 151, Special Issue covering the Nineteenth Symposium on Thermophysical Properties.

(9) Thol, M.; Rutkai, G.; Köster, A.; Dubberke, F. H.; Windmann, T.; Span, R.; Vrabec, J. *Journal of Chemical & Engineering Data* **2016**, *61*, 2580–2595.

(10) Thol, M.; Rutkai, G.; Köster, A.; Miroshnichenko, S.; Wagner, W.; Vrabec, J.; Span, R. *Molecular Physics* **2017**, *115*, 1166–1185.

(11) Martin, M. G.; Siepmann, J. I. *The Journal of Physical Chemistry B* **1998**, *102*, 2569–2577.

(12) Martin, M. G.; Siepmann, J. I. *The Journal of Physical Chemistry B* **1999**, *103*, 4508–4517.

(13) Shah, M. S.; Siepmann, J. I.; Tsapatsis, M. *AIChE Journal* **2017**, *63*, 5098–5110.

(14) Errington, J. R.; Panagiotopoulos, A. Z. *The Journal of Physical Chemistry B* **1999**, *103*, 6314–6322.

(15) Ungerer, P.; Beauvais, C.; Delhommelle, J.; Boutin, A.; Rousseau, B.; Fuchs, A. H. *The Journal of Chemical Physics* **2000**, *112*, 5499–5510.

(16) Bourasseau, E.; Ungerer, P.; Boutin, A.; Fuchs, A. H. *Molecular Simulation* **2002**, *28*, 317–336.

(17) Potoff, J. J.; Bernard-Brunel, D. A. *The Journal of Physical Chemistry B* **2009**, *113*, 14725–14731.

(18) Mick, J. R.; Soroush Barhaghi, M.; Jackman, B.; Schwiebert, L.; Potoff, J. J. *Journal of Chemical & Engineering Data* **2017**, *62*, 1806–1818.

(19) Hemmen, A.; Gross, J. *The Journal of Physical Chemistry B* **2015**, *119*, 11695–11707.

(20) Weidler, D.; Gross, J. *Industrial & Engineering Chemistry Research* **2016**, *55*, 12123–12132.

(21) Stöbener, K.; Klein, P.; Horsch, M.; Kufer, K.; Hasse, H. *Fluid Phase Equilibria* **2016**, *411*, 33 – 42.

(22) Messerly, R. A.; KnottsIV, T. A.; Wilding, W. V. *The Journal of Chemical Physics* **2017**, *146*, 194110.

(23) Abraham, M.; van der Spoel, D.; Lindahl, E.; B.Hess,; the GROMACS development team, GROMACS User Manual version 2018, www.gromacs.org (2018).

(24) Razavi, S. M. Optimization of a Transferable Shifted Force Field for Interfaces and Inhomogenous Fluids using Thermodynamic Integration. M.Sc. thesis, The University of Akron, 2016.

(25) Messerly, R. A.; Shirts, M. R. *Journal of Chemical Theory and Computation* **2018**,

(26) Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids*; Clarendon Press ; Oxford University Press: Oxford England New York, 1987; pp xix, 385 p.

(27) Nieto-Draghi, C.; Bocahut, A.; Creton, B.; Have, P.; Ghoufi, A.; Wender, A.; ; Boutin, A.; Rousseau, B.; Normand, L. *Molecular Simulation* **2008**, *34*, 211–230.

(28) Angelikopoulos, P.; Papadimitriou, C.; Koumoutsakos, P. *The Journal of Chemical Physics* **2012**, *137*, 144103.

(29) Frenkel, M.; Chirico, R. D.; Diky, V.; Yan, X.; Dong, Q.; Muzny, C. *Journal of Chemical Information and Modeling* **2005**, *45*, 816–838.