

Humboldt Research Fellowship

Richard Messerly¹

¹*Thermodynamics Research Center (TRC), National Institute of Standards and Technology (NIST), Boulder, Colorado, 80305, USA*

ABSTRACT

Jadran, I have included outlines for two different proposals. The first proposal is to use more realistic nonbonded potentials to improve the high pressure extrapolation. The second proposal is to move beyond the Lorentz-Berthelot combining rules to improve the prediction of mixture properties. I would appreciate your insight regarding which of these is more likely to be funded and which is more important and/or interesting.

1 Rough draft: Developing more accurate nonbonded potentials for EOS development

1.1 Introduction

As most chemical processes contain mixtures of several molecular species, the design of efficient and reliable chemical processes requires accurate prediction of thermophysical properties over a wide range of temperatures, pressures, and compositions.

Furthermore, as most chemical processes contain mixtures of several species, mixture data are extremely valuable and important for the design of chemical processes. Equations of state over a wide range of temperatures and pressures.

The most chemical processes contain mixtures of several species, mixture data are extremely valuable and important for the design of chemical processes. Unfortunately, mixture data are especially scarce and the reliability is even more dubious than pure component data.

The design of efficient and reliable chemical processes requires accurate prediction of thermophysical properties over a wide range of temperatures and pressures. Fundamental equations of state (FEOS), such as those based on the Helmholtz free energy (e.g. REFPROP), are a powerful approach for estimating pressure, density, temperature ($P\rho T$) behavior and caloric properties, such as internal energy (U) and isochoric/isobaric heat capacities (c_v and c_p , respectively). Unfortunately, most compounds do not have sufficient *reliable* experimental data for a diverse set of thermodynamic properties covering a wide range of $P\rho T$ conditions to develop a highly-accurate FEOS. The lack of experimental data at high temperatures and pressures, especially, is attributed to the inherent safety, cost, and complexity of such experiments.

As most chemical processes contain mixtures of several species, reliable estimation of mixture properties is extremely important. Although the FEOS is capable of predicting mixture properties, typically the deviations are larger than the pure species FEOS. One reason for the deficiency in mixture FEOS models is the scarcity of reliable mixture data. For this reason, applying the hybrid data set approach to mixtures is a potentially promising avenue. Developing reliable force fields and FEOS for pure species is an essential first step before focusing on multi-component systems.

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. For this reason, molecular simulation results at extreme temperatures and pressures can supplement experimental data when developing fundamental equations of state. To improve the performance of this so-called “hybrid data set” approach, it is imperative that the force field be transferable over different $P\rho T$ conditions. In a previous study, we demonstrated the poor transferability to high pressures of the popular united-atom Mie n -6 potential (of which the traditional Lennard-Jones 12-6 is a subclass). Therefore, we propose that alternative potential models be considered for developing FEOS. Specifically, we propose the development of extended Lennard-Jones (ex-LJ) force fields to improve the fundamental equations of state at high pressures.

1.2 Theory

Fundamental equations of state based on the Helmholtz free energy (HFE-FEOS) are expressed as

$$\frac{A}{R_g T} \equiv \alpha = \alpha^{\text{ig}} + \alpha^{\text{r}} \quad (1)$$

where A is the Helmholtz free energy, R_g is the gas constant, T is temperature, α is reduced Helmholtz free energy, α^{ig} and α^{r} are the ideal gas and residual contributions to the reduced Helmholtz free energy. As α^{ig} is typically obtained using first principles (*ab initio*) calculations, the primary focus of FEOS development is modeling α^{r} . The state-of-the-art model for pure species is

$$\alpha^{\text{r}} = \alpha_{\text{Pol}}^{\text{r}} + \alpha_{\text{Exp}}^{\text{r}} + \alpha_{\text{GBS}}^{\text{r}} \quad (2)$$

Note that Equation 2 is semi-empirical and has a large number of fitting parameters (between 50 and 100). To avoid over-fitting, the number of parameters in Equation 2 depends on the amount of experimental data used to train the FEOS. The FEOS predictions can result in large errors when extrapolated to higher temperatures and pressures than the conditions for the training data. Improvement in an FEOS at high temperatures and pressures necessitates additional data for those conditions.

The primary advantage of HFE-FEOS is that all thermodynamic properties are related to some combination of Helmholtz free energy derivatives with respect to temperature and density. The main disadvantage of HFE-FEOS is that large amounts of reliable experimental data are required which cover a wide range of temperatures and pressures.

For example, the following expressions demonstrate the interdependency of several thermodynamic properties:[?]

$$Z = 1 + A_{01}^{\text{dep}} \quad (3)$$

$$\frac{1}{T} \left(\frac{-\partial Z}{\partial (1/T)} \right)_{\rho} = 1 + A_{01}^{\text{dep}} - A_{11}^{\text{dep}} \quad (4)$$

$$\frac{U^{\text{dep}}}{R_g T} = A_{10}^{\text{dep}} \quad (5)$$

$$\frac{1}{R_g} \left(\frac{\partial U^{\text{dep}}}{\partial T} \right)_{\rho} = -A_{20}^{\text{dep}} \quad (6)$$

$$\rho \left(\frac{\partial Z}{\partial \rho} \right)_T = 1 + 2A_{01}^{\text{dep}} + A_{02}^{\text{dep}} \quad (7)$$

where we introduce the notation

$$A_{xy}^{\text{dep}} R_g T \equiv (1/T)^x \rho^y \frac{\partial^{x+y} A^{\text{dep}}}{\partial (1/T)^x \partial \rho^y} \quad (8)$$

The development of fundamental equations of state (FEOS) based on the Helmholtz free energy is by no means novel. However, the use of hybrid data sets consisting of experimental and molecular simulation values has only been implemented in the last five years.

The simulation values that are typically included in hybrid data sets are derivatives of the residual Helmholtz free energy with respect to inverse temperature and/or density:^{?, ?, ?, ?, ?}

$$A_{xy}^r R_g T \equiv (1/T)^x \rho^y \frac{\partial^{x+y} A^r}{\partial (1/T)^x \partial \rho^y} \quad (9)$$

Note that *ms2*,[?] the open-source simulation package developed by Jadran Vrabec's group, already computes A_{xy}^r . Therefore, the infrastructure is already in place to implement the hybrid data set approach at the Technische Universitat Berlin.

We would like to emphasize the advantage of using A_{xy}^{dep} for developing FEOS, as this approach eliminates redundant information found in traditional macroscopic properties.^{?, ?, ?, ?, ?} Furthermore, experimental data are typically measured for properties that relate only to first and second derivatives. By contrast, in principle, molecular simulation provides an avenue for estimating third derivatives and higher order terms, which are valuable when fitting FEOS.

We would like to emphasize the advantage of using A_{xy}^{dep} for developing FEOS, as this approach eliminates redundant information found in traditional macroscopic properties.^{?, ?, ?, ?, ?} For example, the following expressions demonstrate the interdependency of several thermodynamic properties:[?]

$$Z = 1 + A_{01}^{\text{dep}} \quad (10)$$

$$\frac{1}{T} \left(\frac{-\partial Z}{\partial (1/T)} \right)_\rho = 1 + A_{01}^{\text{dep}} - A_{11}^{\text{dep}} \quad (11)$$

$$\frac{U^{\text{dep}}}{R_g T} = A_{10}^{\text{dep}} \quad (12)$$

$$\frac{1}{R_g} \left(\frac{\partial U^{\text{dep}}}{\partial T} \right)_\rho = -A_{20}^{\text{dep}} \quad (13)$$

$$\rho \left(\frac{\partial Z}{\partial \rho} \right)_T = 1 + 2A_{01}^{\text{dep}} + A_{02}^{\text{dep}} \quad (14)$$

The Lennard-Jones 12-6 potential is limited

Several variations of the ex-LJ potential can be proposed, e.g., 12-10-8-6, 14-10-8-6, 14-12-8-6.

For example, the 12-10-8-6 ex-LJ potential is expressed as

$$U_{\text{ex-LJ}} = C_{12} r^{-12} + C_{10} r^{-10} + C_8 r^{-8} + C_6 r^{-6} \quad (15)$$

where C_i is the coefficient for the r^{-i} term. Note that there must be at least one positive and one negative coefficient to account for the repulsive and attractive contributions, respectively. By solving a system equations, the coefficients can be related to the more traditional non-bonded parameters, namely, ϵ , σ , and r_{min} :

$$U_{\text{ex-LJ}}(r_{\text{min}}) = -\epsilon \quad (16)$$

$$U_{\text{ex-LJ}}(\sigma) = 0 \quad (17)$$

$$\frac{\partial U_{\text{ex-LJ}}(r_{\text{min}})}{\partial r} = 0 \quad (18)$$

Also, by defining ϵ , σ , and r_{\min}

The primary deficiency in the Mie λ -6 potential is that the optimal value of λ is greater than 12 for reproducing vapor-liquid equilibria properties. Unfortunately, this conclusion is primarily

Improved extrapolation towards high pressure requires a more physically realistic non-bonded potential.

The extended Lennard-Jones potential

The ex-LJ 12-10-8-6 potential is more flexible than the two-parameter (ϵ and σ) LJ 12-6, and more theoretically justified than the three-parameter Mie n -6. However, it has not been tested as extensively as the LJ 12-6, Mie n -6, and exponential-6 potentials. Although the ex-LJ was proposed over two decades previously, the main reason for the lack of popularity is the additional complexity in parameterizing the ex-LJ potential. Due to the additional model parameters, the traditional approach

Whereas the LJ 12-6, Mie n -6, and expo

For example, the traditional approach to parameterize the Mie n -6 potential is to compare experimental vapor-liquid equilibria data with molecular simulation values for different sets of ϵ , σ , and n . For the ex-LJ potential, this brute-force approach is computationally infeasible. For this reason, we propose an alternative optimization scheme that was recently developed.

higher dimensional parameterization of the ex-LJ is computationally demanding when the force field is parameterized using traditional brute-force methods.

One reason for the lack of popularity is the increased complexity due to the higher dimensional parameterization.

A large amount of experimental data are required to obtain a unique set of non-bonded parameters.

To facilitate parameterization of ex-LJ potentials, we propose the use of Multistate Bennett Acceptance Ratio (MBAR) combined with basis functions (BF). Previous publications demonstrate that MBAR-BF reduces the computational cost by several orders of magnitude compared to direct molecular simulation. The combined MBAR-BF approach estimates ensemble averages for MBAR has been shown to be a reliable approach for reweighting conf has been shown to greatly accelerate force field parameterization

MBAR reweights configurations that are sampled with a reference force field to predict thermodynamic properties for a non-simulated force field. MBAR requires "recalculating" the non-bonded energies and forces for each configuration sampled. Basis functions greatly accelerate the cost of this recalculation step by storing the $\sum r^{-m}$ contributions. Basis functions are amenable to the extended Lennard-Jones potential. For example, the

Basis functions store the contributions from the r^{-m} terms for the nonbonded energies and forces

1.3 Methods

The methodology proposed in this study is to:

1. Develop a FEOS over the temperatures and pressures where reliable experimental data exist
2. Fit the ex-LJ parameters to FEOS derivative Helmholtz free energy properties in range where FEOS is reliable
3. Perform molecular simulations at state points where experimental data are not available
4. Iterate:
 - (a) Refit the FEOS to the hybrid data set

- (b) Re-optimize the force field parameters to match FEOS (using MBAR so that no additional molecular simulations are required)

With this iterative approach, it is possible to ensure that the Helmholtz free energy derivative properties are internally consistent between the FEOS and the ex-LJ force field. The purpose for this iterative approach is to improve the extrapolation of FEOS and the reliability of the force field.

With this iterative approach, it is possible to ensure that the force field derivative properties are internally consistent with those of the FEOS, resulting in better extrapolation for both the EOS and the force field.

As the ex-LJ potential has not been studied extensively, a significant step in this proposal is to develop new combining rules for cross interactions. For example, applying the traditional Lorentz-Berthelot combining rules allows for estimation of ϵ_{ij} , σ_{ij} , and/or $r_{\min,ij}$ from ϵ_{ii} , σ_{ii} , and/or $r_{\min,ii}$. Even in pure species systems, cross interactions are typically implemented to simplify the parameterization and to avoid non-physical parameters. The cross interaction parameters play a key role in mixture property estimates.

For example, the traditional Lorentz-Berthelot combining rules allow for estimation of ϵ_{ij} , σ_{ij} , and/or $r_{\min,ij}$ from the .

The focus of this research is pure species properties

Important to develop new combining rules for cross interactions of ex-LJ. Additional binary interaction parameters improve agreement with mixture properties

Although the focus of this research is developing highly accurate FEOS for pure species, it should be noted that reliable estimation of mixture properties is an extremely important area.

2 Outline: Developing more accurate nonbonded potentials for EOS development

1. Background to establish need for hybrid data sets
 - (a) Design of efficient and reliable chemical processes requires accurate equations of state over a wide range of temperatures and pressures
 - (b) The quantity and quality of experimental data is insufficient for most compounds
 - (c) Molecular simulation is an ideal approach for overcoming the data deficiency at extreme temperatures and pressures
 - (d) Improved EOS are developed by supplementing experimental data with simulation results
2. Reliability of hybrid data set approach depends on accuracy of force field, specifically, how well the model extrapolates to extreme temperatures and pressures
 - (a) Recent work has demonstrated that the united-atom Lennard-Jones n -6 force field does not extrapolate from vapor-liquid equilibria conditions to elevated pressures
 - (b) Improved extrapolation requires a more physically realistic nonbonded potential
 - (c) For this reason, we propose the use of an extended Lennard-Jones (ex-LJ, 12-10-8-6) potential
3. The ex-LJ 12-10-8-6 potential is more flexible than the LJ 12-6, and more theoretically justified than the Mie n -6
4. Although the ex-LJ was proposed over two decades previously, it has not been tested as extensively as the LJ 12-6, Mie n -6, and exponential-6

5. One reason for the lack of popularity is that the higher dimensional parameterization of the ex-LJ is computationally infeasible when implementing traditional iterative force field optimization approaches
 - (a) Multistate Bennett Acceptance Ratio (MBAR) combined with basis functions (BF) has been shown to greatly accelerate force field parameterization
 - (b) MBAR reweights configurations sampled with a reference force field to predict thermodynamic properties for a non-simulated force field
 - (c) Basis functions store the contributions from the r^{-m} terms for the nonbonded energies and forces
6. The methodology proposed in this study is to:
 - (a) Develop an EOS over the temperatures and pressures where reliable experimental data exist
 - (b) Fit the ex-LJ parameters to EOS derivative Helmholtz energy properties in range where EOS is reliable
 - (c) Perform molecular simulations at state points where experimental data are not available
 - (d) Iterate:
 - i. Refit the EOS to the hybrid data set
 - ii. Reoptimize the force field parameters to match EOS (using MBAR so that no additional molecular simulations are required)
7. With this iterative approach, it is possible to ensure that the force field derivative properties are internally consistent with those of the EOS, resulting in better extrapolation for both the EOS and the force field
8. Important to develop new combining rules for cross interactions of ex-LJ
9. Additional binary interaction parameters improve agreement with mixture properties

3 Justification for funding

1. Why is this important?
 - (a) Reliable estimates of thermophysical properties have significant practical industrial applications
2. Why me?
 - (a) Expertise in force field parameterization and, specifically, the MBAR-BF method
 - (b) Expertise in both molecular dynamics and Monte Carlo simulations
3. Novel aspects of this research to expand my expertise:
 - (a) Fitting fundamental equations of state
 - (b) Residual helmholtz energy derivatives from molecular simulation
 - (c) Hybrid data set approach
4. Why Jadran?
 - (a) Dr. Vrabec has helped pioneer the FEOS hybrid data set approach

- (b) Substantial computational resources at disposal
- (c) Strong research group and collaborators
- (d) Experience developing molecular simulation software

Importance sampling is a commonly used statistical technique for computing averages of properties for one model by reweighting configurations sampled with another model based on the ratio of probabilities for the two models. In chemistry and chemical physics, importance sampling from one or a set of simulations to another set of simulation conditions can be implemented using the MBAR algorithm.^{2,7} With MBAR the expectation $\langle O(\theta) \rangle$ for force field (θ) of any given observable (O) can be expressed as:

$$\langle O(\theta) \rangle = \sum_{n=1}^N O(x_n; \theta) W_n(\theta) \quad (19)$$

where x_n are configurations sampled from one or more reference force fields (θ_{ref}), $O(x_n; \theta)$ is the observable value using force field θ with configurations x_n , and $W_n(\theta)$ is the weight of the n^{th} configuration using force field θ , calculated by using:

$$W_n(\theta) = \frac{\exp[\hat{f}(\theta) - u(x_n; \theta)]}{\sum_{k=1}^K N_k \exp[\hat{f}(\theta_{\text{ref},k}) - u(x_n; \theta_{\text{ref},k})]} \quad (20)$$

where the reduced free energies ($\hat{f}(\theta)$) are calculated with:

$$\hat{f}(\theta) = -\ln \sum_{n=1}^N \frac{\exp[-u(x_n; \theta)]}{\sum_{k=1}^K N_k \exp[\hat{f}(\theta_{\text{ref},k}) - u(x_n; \theta_{\text{ref},k})]} \quad (21)$$

where K is the number of reference force fields, $N = \sum_k N_k$ is the total number of snapshots for all K reference force fields, N_k are the total number of snapshots from the k^{th} reference force field, $\theta_{\text{ref},k}$ is the k^{th} reference (i.e. simulated) force field, and $u(x_n; \theta) = \beta U(x_n; \theta)$ is the reduced potential energy evaluated with θ for configuration x_n where $\beta = \frac{1}{k_B T}$ and k_B is the Boltzmann constant.

Note that $\hat{f}(\theta_{\text{ref},k})$ is required to evaluate the denominator of Equations 20-21. The values for $\hat{f}(\theta_{\text{ref},k})$ are obtained by solving a system of K equations for self-consistency. Specifically, an initial guess for $\hat{f}(\theta_{\text{ref},k})$ is used to evaluate Equation 21 with $\theta = \theta_{\text{ref},k}$ to obtain updated values of $\hat{f}(\theta_{\text{ref},k})$. This process is repeated until the values for $\hat{f}(\theta_{\text{ref},k})$ converge to within a desired tolerance. Although solving the MBAR system of equations for self-consistency may require several iterations, fortunately, once this process has been performed $\hat{f}(\theta)$ (for an arbitrary θ) is evaluated without further iteration.

For the specific case of predicting U^{dep} and Z for a non-sampled force field, expressed by the set of force field parameters θ , the MBAR-based estimators for the departure internal energy and compressibility can be written as:

$$\langle U^{\text{dep}}(\theta) \rangle = \sum_{n=1}^N U^{\text{dep}}(x_n; \theta) W_n(\theta) \quad (22)$$

$$\langle Z(\theta) \rangle = \sum_{n=1}^N Z(x_n; \theta) W_n(\theta) \quad (23)$$

where the energies and forces are computed using force field θ for each configuration (x_n) to determine $U^{\text{dep}}(x_n; \theta)$ and $Z(x_n; \theta)$ (from the virial pressure²), respectively, while the weights

$(W_n(\theta))$ are again calculated using Equations 20-21. The performance of MBAR depends strongly on good phase space overlap, meaning that the configurations sampled by the reference force field(s) must represent a significant portion of the “true” configurations that the non-simulated force field would sample.⁷ If the configurational overlap is small, the MBAR estimates are often dominated by a few configurations, which are likely not representative of the ensemble that would be generated by direct simulation of force field θ . The amount of overlap can be quantified by the number of effective samples (N_{eff}),⁷ using Kish’s formula:

$$N_{\text{eff}} = \frac{(\sum_n W_n)^2}{\sum_n W_n^2} \quad (24)$$

which reduces to $N_{\text{eff}} = (\sum_n W_n^2)^{-1}$ when the weights are normalized. This has the property that when the weights are equal, $N_{\text{eff}} = N$, when all but one weight is negligible, $N_{\text{eff}} \approx 1$, and behaves appropriately for intermediate cases. In the case of poor overlap ($N_{\text{eff}} \approx 1$), the predicted values of MBAR will demonstrate a strong bias and the uncertainties will likely be underestimated by the MBAR covariance matrix.