Advancing the hybrid data set approach towards reliable mixture properties: Going beyond the traditional Lennard-Jones potential

The understanding of natural mechanisms and design of efficient and reliable chemical processes rests on the accurate knowledge of thermophysical properties of fluids over a wide range of temperatures and pressures. Fundamental equations of state (FEOS), such as those based on the Helmholtz energy (A), are a powerful approach for consistently representing the 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) of pure species and mixtures. The National Institute of Standards and Technology (NIST) is the global leader in FEOS development. During my postdoctoral associateship at NIST, I have been fortunate to collaborate with the developers of the NIST FEOS, Reference Fluid Properties (REFPROP [1]).

In general, FEOS separate the Helmholtz energy into an ideal gas (A^ig) and a residual (A^r) contribution. As A^ig is typically obtained using first principles (*ab initio*) calculations, the primary focus of FEOS development is modeling A^r. State-of-the-art FEOS, such as REFPROP, utilize a semi-empirical model for A^r with between 50 and 100 non-linear fitting parameters. The number of fitting parameters and, thence, the FEOS accuracy is determined by the quality, quantity, and diversity of experimental data. When an FEOS correlation is reliable over the entire fluid region of technological interest, it is considered to be of "technical accuracy."

Since most fluids are mixtures of several molecular species, reliable estimates of mixture properties are essential. The development of mixture FEOS is, thus, a key endeavor at NIST and several sessions were dedicated to this topic at the 20th Symposium on Thermophysical Properties organized by NIST in 2018. The primary challenge is that mixture FEOS require additional fitting parameters for A^r, while experimental mixture data are scarce (due to the additional composition variable) and of questionable quality. Furthermore, any deficiencies in the pure species FEOS propagate when predicting mixture properties. Therefore, the ability to develop "technical accuracy" mixture FEOS necessitates "technical accuracy" pure species FEOS. For this reason, the discussion that follows focuses on pure species.

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 "technical accuracy" FEOS. For example, REFPROP covers less than 100 compounds with "technical accuracy." Due to the large amount of fitting parameters, FEOS predictions can result in substantial errors when extrapolated to higher temperatures and pressures than those of the training data. Improvement in an FEOS necessitates additional data at these extreme temperatures and pressures. The lack of experimental data at these conditions is attributed to the inherent safety, cost, and complexity of such experiments (e.g., chemical instability above the thermal decomposition temperature).

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, the host, Vrabec, and coworkers proposed the use of "hybrid data sets" consisting of both available experimental data and molecular simulation results at extreme temperatures and pressures [2]. Together with collaborators, Vrabec

has demonstrated for several compounds that hybrid data sets extend the range of applicability for the FEOS [2, 3, 4, 5, 6].

The simulation values that are included in hybrid data sets are derivatives of the residual Helmholtz energy with respect to inverse temperature and/or density

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

where R_g is the gas constant. We would like to emphasize the advantage of using A_{xy}^{r} for developing FEOS, as this choice eliminates redundant information found in traditional macroscopic properties [2, 7, 8, 9, 10, 11]. Furthermore, experimental data are typically measured for properties that relate only to first and second derivatives, whereas, in principle, molecular simulation provides an avenue for estimating higher order derivatives, which could be invaluable when fitting FEOS.

The primary limitation of the hybrid data set approach is that most force fields are not "transferable" over a wide range of $P\rho T$ conditions. This deficiency causes the force field to be inconsistent with the FEOS. For example, in a recent study [12], we demonstrated poor transferability to high pressures of the popular Mie n-6 potential for normal and branched alkanes. Specifically, as the Mie n-6 potential requires $n \geq 14$ to accurately predict vapor-liquid equilibria (VLE) properties, it is overly-repulsive at short distances, resulting in significant deviations at high densities/pressures. Therefore, Mie n-6 force fields (of which the traditional Lennard-Jones 12-6 is a subclass) are ill-suited for the hybrid data set approach.

The question that this study will address is, can a pairwise additive force field have the same "technical accuracy" as the FEOS? As thermodynamic properties are extremely sensitive to intermolecular interactions, **we propose that more theoretical and flexible non-bonded potential functions be considered when developing FEOS.** For example, two-body potentials developed from *ab initio* values typically utilize more than ten fitting parameters because of the high information content in *ab initio* energy calculations performed at different interatomic distances (see Equation 6 of Reference 13).

For nearly half a century the Lennard-Jones (LJ) 12-6 potential has inundated the molecular simulation literature (the popular choice of $n = 12$ has no theoretical basis and is a historical artifact based primarily on computational reasons that are no longer significant). Only in the past decade has the Mie n-6 potential received considerable attention. However, due to inadequacies in the Mie n-6 potential at high pressures, **we propose the development of extended Lennard-Jones (ex-LJ) force fields for the hybrid data set approach.**

The general expression for the extended Lennard-Jones non-bonded potential is

$$u_{\mathrm{nb,ex\text{-}LJ}}(\mathbf{C}) = \sum_{m=6,8\ldots} C_m r^{-m} \tag{2}$$

where m are integer (typically even) values and the parameter set, $\mathbf{C}$, consists of the C_m coefficients corresponding to the r^{-m} terms. Note that the traditional Lennard-Jones 12-6 potential is obtained if $C_{12} = 4\varepsilon\sigma^{12}$ and $C_6 = -4\varepsilon\sigma^6$ (where σ and ε are the Lennard-Jones size and energy parameters, respectively), while all other C_m values are zero. The attractive terms (negative C_m) are derived rigorously from London dispersion forces [14]. By contrast, the repulsive terms (positive C_m) are strictly empirical

and, if necessary, can be replaced by more theoretical functions (e.g., $\exp(-r)$ [13, 15]).

The ex-LJ 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, particularly when $n \gg 12$. However, it has not been tested as extensively as the LJ 12-6, Mie n-6, and exponential-6 potentials. By demonstrating significant improvement at high pressures, **the potential impact of this research would be to initiate a paradigm shift in non-bonded potentials.**

The question remains, when enough non-zero terms are included in Equation 2, can an ex-LJ force field fit higher order derivatives of the Helmholtz energy over the entire fluid region of technological interest? The development of "technical accuracy" ex-LJ force fields will allow for prediction of other properties not available from the FEOS, such as transport properties (e.g., self-diffusivity, shear viscosity) and structural/kinetic phenomena, as well as various mixture properties.

As the ex-LJ has received relatively little attention in the literature, some additional questions naturally arise. For example, since most *ab initio* based two-body potentials utilize r^{-6}, r^{-8}, and r^{-10} terms for attractive interactions, should the C_6, C_8, and C_{10} coefficients necessarily be negative? How many non-zero terms in Equation 2 should be included such that the model is sufficiently flexible but not over-fit? Similarly, which terms provide the greatest improvement in the force field? Which combining rules (e.g., Lorentz-Berthelot) should be used for cross interactions?

Combining rules reduce the number of force field parameters that are optimized simultaneously for compounds with multiple interaction site types. More importantly, combining rules are essential for performing simulations of mixtures, which is a key motivation for developing "technical accuracy" force fields for pure species. While numerous combining rules are proposed in the literature for the Lennard-Jones parameters (cf., Reference 16), combining rules for C_m are less straightforward and will likely require innovative formulations.

Although the ex-LJ was proposed over three decades ago [17], the main reason for the lack of popularity is the additional complexity in parameterizing the ex-LJ potential when several C_m terms are non-zero. This presents another essential question, what optimization method is best suited for parameterizing the ex-LJ potential? The development of an optimization scheme will **allow future researchers to parameterize "technical accuracy" ex-LJ force fields.**

In general, *ab initio* based two-body potentials do not perform well in liquid and supercritical phases and are, therefore, ill-suited for developing "technical accuracy" force fields. Instead, effective non-bonded potentials (which indirectly account for three-body and higher-body interactions) are parameterized empirically to macroscopic condensed phase properties. For example, previous hybrid data set studies optimized the non-bonded parameters with VLE data, e.g., saturated liquid density and saturated vapor pressure. However, it is unlikely that VLE properties alone can provide a unique set of parameters for a highly flexible potential, such as the ex-LJ with more than three non-zero C_m terms. By contrast, derivatives of Helmholtz energy (A^r_{xy}) provide high information content regarding the non-bonded potential. Unfortunately, reliable A^r_{xy} values require an accurate FEOS, which is not available *a priori*. For this reason, **we propose a *novel* iterative hybrid data set approach** (see Figure 1).

The iterative approach presented in Figure 1 ensures that the Helmholtz energy derivative properties are internally consistent between the FEOS and the ex-LJ force field. The aim is to improve the extrapolation of the FEOS and the force field transfer-
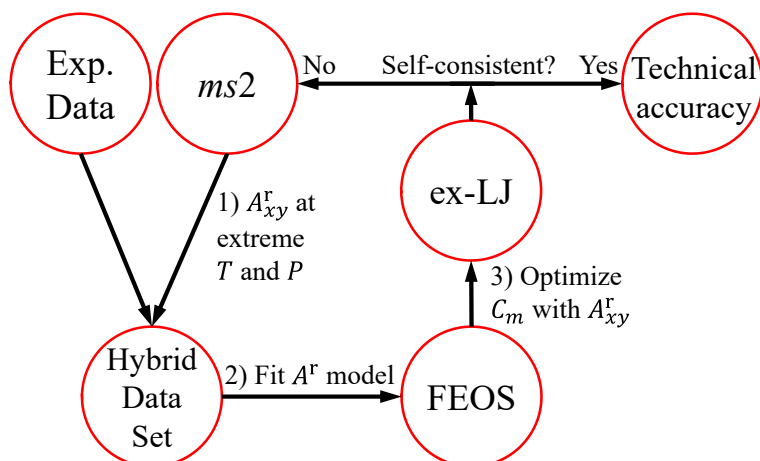
Figure 1: Iterative hybrid data set approach for generating self-consistent FEOS and ex-LJ force field of "technical accuracy."

ability. As the FEOS and ex-LJ force field become more self-consistent with each successive iteration, it is possible to increase the number of non-zero C_m terms in Step 3), although this may necessitate including higher order derivatives (i.e., A^r_{xy} for $x + y > 2$).

The infrastructure is already in place to implement Step 1) at the host institute, Technische Universität Berlin. Specifically, Vrabec's group developed $ms2$ [18], a highly optimized and parallelized code that is capable of performing both MC and MD simulations. More importantly, $ms2$ is the only open-source simulation package that computes A^r_{xy} for various values of x and y. In addition, Vrabec's group has already automated the entire process of simulating the necessary $P\rho T$ state space and processing the A^r_{xy} results with minimal human interaction. With access to tens of thousands of cores on national supercomputers, Vrabec's group is capable of generating all the required molecular simulation results for a given force field (Step 1)) in just a few hours.

Step 3) of this algorithm is the computational bottleneck when direct molecular simulations are performed for each re-parameterization of the ex-LJ force field. In fact, the traditional brute-force trial-and-error optimization approach is not computationally feasible for more than three non-zero C_m terms. To facilitate parameterization of ex-LJ potentials with A^r_{xy}, we propose the use of Multistate Bennett Acceptance Ratio (MBAR) combined with basis functions (Φ). **My previous publications demonstrated that MBAR-Φ reduces the computational cost to estimate ensemble averages by several orders of magnitude compared to direct molecular simulation** [12, 19]. Therefore, utilizing MBAR-Φ in Step 3) is essential for this algorithm to be computationally tractable. In addition, MBAR-Φ can be applied in Step 1) to eliminate the need to re-simulate the high temperature and pressure state points for each iteration. Due to the essential role of MBAR-Φ in this algorithm, we now present a brief overview of this method.

MBAR is a statistical method that reweights configurations sampled with a reference force field(s) to predict ensemble averages for a non-simulated force field [19, 20]. For example, derivatives of the Helmholtz energy for parameter set **C** are estimated

4

according to

$$A_{xy}^{\mathrm{r}}(\mathbf{C}) = F_i[\langle U(q_{\mathrm{ref}}, \mathbf{C})^i \rangle_{\mathrm{MBAR}}] + F_j[\langle U(q_{\mathrm{ref}}, \mathbf{C})^j \rangle_{\mathrm{MBAR}}] \tag{3}$$

where q_{ref} are configurations sampled using the reference force field(s) ($\mathbf{C}_{\mathrm{ref}}$), $\langle \rangle_{\mathrm{MBAR}}$ are ensemble averages estimated using MBAR (see Equations 9 to 11 of Reference 19), and F_i and F_j are functionals that depend on different powers of the internal energy (see Equations 27 and 30 of Reference 21). Since MBAR has never been implemented to estimate A_{xy}^{r}, an important step in this project is to develop best practices for evaluating Equation 3 (e.g., the optimal number of snapshots in q_{ref}) and to assess its range of reliability (e.g., the accuracy for large differences between $\mathbf{C}_{\mathrm{ref}}$ and $\mathbf{C}$).

While MBAR reweighting is orders of magnitude faster than performing a direct molecular simulation, MBAR requires "recalculating" the non-bonded energies for each configuration sampled, q_{ref}. The total non-bonded internal energy ($U_{\mathrm{nb,total}}$) is computed from $\sum_{i=1}^{N_{\mathrm{sites}}-1} \sum_{j=i+1}^{N_{\mathrm{sites}}} \sum_m C_{m,ij} r_{ij}^{-m}$, where N_{sites} is the number of interacting sites in the system, $C_{m,ij}$ is the C_m term for the ij interaction and r_{ij} is the intermolecular distance between sites i and j. Rather than storing the configurations of all N molecules, basis functions store the $\sum_{i=1}^{N_{\mathrm{sites}}-1} \sum_{j=i+1}^{N_{\mathrm{sites}}} r_{ij}^{-m}$ contributions for different values of m. Due to the linear relationship between $U_{\mathrm{nb,total}}$ and Equation 2, the ex-LJ potential is amenable to basis functions and, therefore, recomputing $U_{\mathrm{nb,total}}$ for a perturbed set of $C_{m,ij}$ parameters requires simple and *fast* matrix multiplication. For example, basis functions reduce the cost of recalculating $U_{\mathrm{nb,total}}$ by several orders of magnitude compared to the Gromacs "rerun" function, which is already highly optimized [22]. Since recomputing $U_{\mathrm{nb,total}}$ with basis functions does not depend on the system size, the computational benefit of basis functions becomes more pronounced as N_{sites} increases, i.e., more/larger molecules.

In summary, MBAR-Φ permits rapid estimation of A_{xy}^{r} for non-simulated extended Lennard-Jones potentials. This renders the iterative hybrid data set approach computationally feasible by removing the need for performing thousands of molecular simulations to re-parameterize the force field. **My expertise with MBAR-Φ, Vrabec's hybrid data set method and computational resources, and our strong connections with NIST and other developers of FEOS are essential for accomplishing the goals of this research, namely, to develop self-consistent FEOS and force fields of "technical accuracy."**

Five deliverables are expected from this project:

1. "Technical accuracy" FEOS for each molecule studied

2. Force fields of "technical accuracy"

3. Increased theoretical understanding of non-bonded potentials

4. Combining rules for extended Lennard-Jones potential

5. Infrastructure for rapid force field parameterization using MBAR-Φ and Helmholtz energy derivatives

References

[1] Lemmon *et al.* National Institute of Standards and Technology, 2018.

[2] Rutkai *et al. The Journal of Chemical Physics*, 139(4):041102, 2013.

[3] Thol *et al. Fluid Phase Equilibria*, 418:133 – 151, 2016.

[4] Thol *et al. Journal of Chemical & Engineering Data*, 61(7):2580–2595, 2016.

[5] Thol *et al. Molecular Physics*, 115(9-12):1166–1185, 2017.

[6] Thol *et al. Chemical Engineering Science*, 121:87 – 99, 2015.

[7] Thol *et al. Journal of Physical and Chemical Reference Data*, 45(2):023101, 2016.

[8] Thol *et al. International Journal of Thermophysics*, 36(1):25–43, Jan 2015.

[9] Rutkai *et al. Molecular Physics*, 115(9-12):1104–1121, 2017.

[10] Lustig *et al. Molecular Physics*, 113(9-10):910–931, 2015.

[11] Rutkai *et al. Journal of Chemical & Engineering Data*, 60(10):2895–2905, 2015.

[12] Messerly *et al. Journal of Chemical Physics*, 2018. Pending publication.

[13] Hellmann *et al. The Journal of Chemical Physics*, 147(3):034304, 2017.

[14] Anthony Stone. The theory of Intermolecular Forces, 2nd edition. 54, 07 2013.

[15] Przybytek *et al. Physical Review Letters*, 119:123401, Sep 2017.

[16] Schnabel *et al. Journal of Molecular Liquids*, 135(1):170 – 178, 2007.

[17] Kalos *et al. Canadian Journal of Chemistry*, 50(6):892–896, 1972.

[18] Rutkai *et al. Computer Physics Communications*, 221:343 – 351, 2017.

[19] Messerly *et al. Journal of Chemical Theory and Computation*, 14(6):3144–3162, 2018.

[20] Shirts and Chodera. *The Journal of Chemical Physics*, 129:124105, 2008.

[21] Rolf Lustig. *Molecular Physics*, 110(24):3041–3052, 2012.

[22] GROMACS development team. *GROMACS User Manual version 2018*.