

in the context of force fields, "transferability" typically relates to "from one species to the other"
 → maybe: "do not lead to convincing results"?

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

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

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. *applicability at?*

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 equilibrium (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 flexible and physically realistic non-bonded potential functions be considered when developing FEOS.** For example, *ab initio* based two-body potentials typically utilize more than ten fitting parameters (see Equation 6 of Reference 13).

great!

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

$$u_{\text{nb,ex-LJ}}(C) = \sum_{m=6,8,\dots} C_m r^{-m} \quad (2)$$

where m are integer (typically even) values and the parameter set, 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 $C_{12} = 4\epsilon\sigma^{12}$ and $C_6 = -4\epsilon\sigma^6$ (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 a sufficient number of non-zero terms is 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/non-equilibrium 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 [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.

Ab initio energy calculations performed at different intermolecular distances have high information content for the optimization of non-bonded interactions. Unfortunately, *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. Accounting for three-body interactions is an arduous task with *ab initio* methods, especially for multiple-atom molecules, and provides only marginal improvement. Instead, accurate prediction of condensed phase properties is often achieved by developing effective non-bonded parameters (which indirectly account for three-body and higher-body interactions).

For example, previous hybrid data set studies optimized the non-bonded parameters to 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_{xy}^r) provide high information content regarding the non-bonded potential. Unfortunately, reliable A_{xy}^r 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).

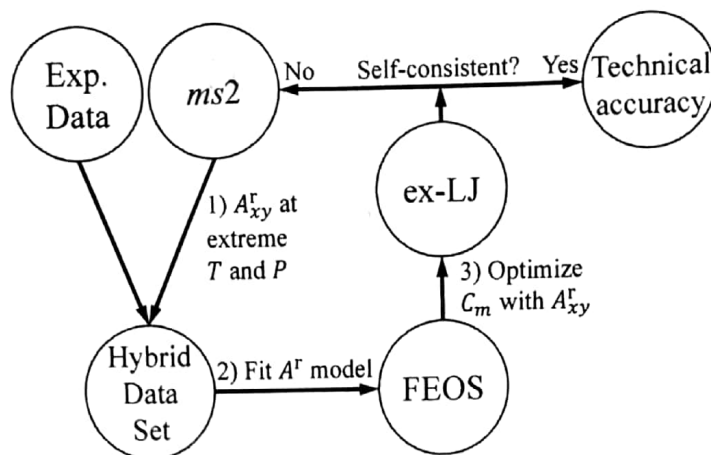


Figure 1: Iterative hybrid data set approach for generating self-consistent FEOS and ex-LJ force field of "technical accuracy."

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 transferability. 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_{xy}^r for $x+y > 2$). see above

The infrastructure is already in place to implement Steps 1) and 2) 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_{xy}^r 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_{xy}^r 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. Vrabec also has access to sophisticated constrained non-linear optimization schemes, which are commonly used for FEOS fitting (Step 2). ?

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_{xy}^r , 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 the proposed approach, 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 \mathbf{C} are estimated according to

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

where q_{ref} are configurations sampled using the reference force field(s) (\mathbf{C}_{ref}), $\langle \rangle_{\text{MBAR}}$ 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}_{ref} and \mathbf{C}).

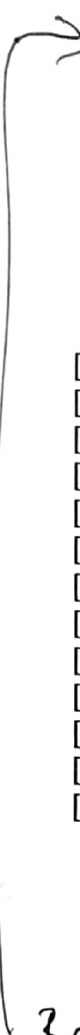
While MBAR reweighting is orders of magnitude faster than performing a direct molecular simulation, MBAR requires "recalculating" the non-bonded energies for each configuration sampled, q_{ref} . The total non-bonded internal energy ($U_{\text{nb,total}}$) is computed from $\sum_{i=1}^{N_{\text{sites}}-1} \sum_{j=i+1}^{N_{\text{sites}}} \sum_m C_{m,ij} r_{ij}^{-m}$, 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 $\sum_{i=1}^{N_{\text{sites}}-1} \sum_{j=i+1}^{N_{\text{sites}}} r_{ij}^{-m}$ contributions for different values of m . Due to the linear relationship between $U_{\text{nb,total}}$ and Equation 2, the ex-LJ potential is amenable to basis functions and, therefore, recomputing $U_{\text{nb,total}}$ for a perturbed set of $C_{m,ij}$ parameters requires simple and fast matrix multiplication. For example, basis functions reduce the cost of recalculating $U_{\text{nb,total}}$ by several orders of magnitude compared to the Gromacs "rerun" function, which is already highly optimized [22]. Since recomputing $U_{\text{nb,total}}$ 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. *only?* *Consequently,* *for?*

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."** The FEOS and force fields (along with the appropriate ex-LJ combining rules) will permit reliable prediction of mixture properties and will, thus, advance the field of mixture FEOS development.

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 A_{xy}^r

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, 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] Stone. *The Theory of Intermolecular Forces*, 2nd edition. 54, 2013.
 - [15] Przybytek *et al.* *Physical Review Letters*, 119:123401, 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] Lustig. *Molecular Physics*, 110(24):3041–3052, 2012.
 - [22] GROMACS development team. *GROMACS User Manual*, version 2018.

Because this is the only reference that cannot be identified because of its generality, you should give its title, or even better, the weblink.

important for physicists and chemists?

Short abstract

Online description

The short abstract of your research proposal should be worded in a manner that is appropriate for non-specialist academics and should summarize the main goals and contents of the research you plan to do in Germany. Please do not use abbreviations without explaining them.

understanding natural phenomena

My text

Reliable estimates of thermophysical properties are essential for designing efficient ~~and reliable~~ technical processes. Fundamental equations of state (FEOS) based on the Helmholtz energy allow for prediction of pressure, density, temperature behavior as well as energetic properties, e.g., heat capacities. Unfortunately, most molecular species (and, to a greater extent, mixtures) do not have enough reliable experimental data to fit the large number of FEOS parameters. In this case, molecular simulation can supplement experimental data at state points where reliable data are scarce, typically at high temperatures and pressures.

much?

The primary limitation of this so-called "hybrid data set" approach is the accuracy of the force field used in the molecular simulation. Specifically, most force fields perform well for vapor-liquid equilibrium properties but extrapolate poorly to high pressures. As thermophysical properties are highly sensitive to the non-bonded interactions, we propose using the extended Lennard-Jones (ex-LJ) potential, which is significantly more flexible than the traditional Lennard-Jones 12-6 potential. Furthermore, we propose an iterative hybrid data set approach, where the ex-LJ parameters are re-optimized after each iteration to ensure self-consistency between the FEOS and the force field.

To reduce the computational cost of this iterative approach, we will implement Multistate Bennett Acceptance Ratio (MBAR) combined with basis functions. In my previous work, I demonstrated that MBAR with basis functions yields extremely fast and reliable estimates of thermophysical property values for any force field parameter set, without performing direct molecular simulation. My expertise with MBAR and basis functions, the host's simulation infrastructure and methods, and our close collaborations with experts in FEOS development are essential for the success of this project.

additional

NUM3

What do you think will be the impact of your research on the further development of your academic profile?

I intend on making the proposed methodology, "iterative hybrid data sets," the keystone of my future research. It would be impossible to develop equations of state and force fields for every compound of interest during this two-year fellowship. Therefore, I will continue to implement this approach for years to come with additional molecular species.

My long-term career path is to become a professor, but with a strong emphasis on "industrially relevant" research. The proposed research is exemplary of this as it utilizes state-of-the-art scientific/academic methods but with the ultimate impact found in industry. The

primary benefit of this research is that it allows me to pursue a career path in academia, industry, or a government agency.

Working with Vrabec, a pioneer in hybrid data sets, will greatly accelerate and deepen my understanding of this approach. Furthermore, I will explore the details and better appreciate the challenges of fitting equations of state with high-dimensional non-linear models. This skill will be invaluable throughout my career, regardless of the path I pursue. I will learn Fortran 90, an extremely valuable coding language, as I contribute to the molecular simulation package developed by Vrabec's group, *ms2*. By working directly with the developers of *ms2*, I also expect that my coding practices will improve.

As my previous research groups were relatively small, joining Vrabec's group is a great opportunity for me to see firsthand how larger research groups function. Under Vrabec's tutelage, I will develop my own style for teaching, mentoring, managing several projects, and performing research. Being at a German university fosters both a diverse environment and new global connections.

In brief, there are four key facets in which the proposed research will significantly impact the development of my academic profile:

1. Expertise/skills
2. Leadership
3. Diversification
4. Networking

You should add one sentence
on US \leftrightarrow Europe, cultural

A.
✓

List of the selected key publications of Richard Messerly

1. **Richard A. Messerly**, S. Mostafa Razavi, and Michael R. Shirts. "Configuration-sampling-based surrogate models for rapid parameterization of non-bonded interactions." *Journal of Chemical Theory and Computation*. 14 (6), 3144-3162, 2018.

This was my first publication during my postdoctoral associateship at the National Institute of Standards and Technology (NIST). Furthermore, this publication marked the first time that my co-authors were collaborators outside of my own institution. Furthermore, Michael R. Shirts is an extremely well-published young professor and S. Mostafa Razavi is a graduate student for J. Richard Elliott, an established expert in the field. Thus, this was a landmark publication that helped to proliferate my name throughout the molecular simulation community. With respect to the actual content, this publication compared Michael Shirts's existing method (MBAR) with a novel approach I developed (PCFR). The comparison demonstrated that these two methods are complementary and should be implemented together. Furthermore, I demonstrated how to merge these two methods with an approach developed by J. Richard Elliott (ITIC). **Personal contribution: 75%.**

2. **Richard A. Messerly**, Thomas A. Knotts IV, and W. Vincent Wilding. "Uncertainty quantification and propagation of errors of the Lennard-Jones 12-6 parameters for *n*-alkanes." *The Journal of Chemical Physics*. 146, 194110, 1-16, 2017.

This publication was the culmination of my dissertation and led directly to my postdoctoral position with the Thermodynamics Research Center (TRC) at NIST. The main topic is uncertainty quantification (UQ) of thermophysical properties with molecular simulation, which is a strong interest at NIST. After presenting this research, the TRC group leader strongly encouraged me to apply for a National Research Council (NRC) postdoctoral position in his group, which was a pivotal step in developing my academic profile. Furthermore, this publication played ~~a~~ a key role in joining the Open Force Field initiative, since this group utilizes Bayesian UQ methods as part of force field development. As a participant in this initiative, I have created strong connections and learned firsthand from a diverse set of experts in the field of molecular simulation. **Personal contribution: 90%.**

3. **Richard A. Messerly**, Michael R. Shirts, and Andrei F. Kazakov. "Uncertainty quantification confirms unreliable extrapolation toward high pressures for united-atom Mie λ -6 force field." *The Journal of Chemical Physics*. (publisher's acknowledgment of receipt enclosed)

(, manuscript ~~attach~~
in the supplementary material)