## Description of proposal guidelines on Humboldt website

The current state of research should first be briefly described and supported by approximately five revelant publications from the research area (one page max).

The outline should focus on a clear description of the questions you intend to address in your research, their originality and significance for the advancement of the research field (approx. two pages).

Furthermore, the academic methods to be used to achieve these goals should be clearly described and referenced, if appropriate (approx. two pages).

The research outline should comprise approximately five pages in total (including references). Should you significantly exceed this length, you may be asked to cut it down to approximately five pages.

For the purposes of evaluation it must be clearly demonstrated that you yourself have drawn up the main contents independently and agreed them beforehand with your host. Any contents contributed by the host institute must be attributed accordingly.

## Rough draft: Self-consistent fundamental equations of state and force fields of "technical accuracy"

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 ($A$), 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) for pure species and mixtures. In general, FEOS separate the Helmholtz free 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 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." 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. Mixture FEOS often require additional fitting parameters although experimental data are even more scarce and of questionable quality. 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 at high temperatures and pressures necessitates additional data for those conditions. The lack of experimental data at high temperatures and pressures, especially, is attributed to the inherent safety, cost, and complexity of such experiments (e.g., chemical instability above the thermal decomposition temperature, ca., 650 K).

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, Reference 1 proposed the use of "hybrid data sets" consisting of both available experimental data and molecular simulation results at extreme temperatures and pressures. The host and collaborators have demonstrated for several compounds that this hybrid data set extends the range of applicability for the FEOS.[1–5] The simulation values that are included in hybrid data sets are derivatives of the residual Helmholtz free energy with respect to inverse

temperature and/or density :

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

where $R_g$ is the gas constant. We would like to emphasize the advantage of using $A^{\mathrm{dep}}_{xy}$ for developing FEOS, as this choice eliminates redundant information found in traditional macroscopic properties.[1,6–10] 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. Note that *ms2*,[11] the open-source simulation package developed by Jadran Vrabec's group, computes $A^{\mathrm{r}}_{xy}$. Therefore, the infrastructure is already in place to implement the hybrid data set approach at the host institute, Technische Universitat Berlin.

The primary limitation of this approach is that most force fields are not "transferable" over a wide range of $P\rho T$ conditions. For example, 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) for normal and branched alkanes. This deficiency in the force field causes the FEOS to be inconsistent with the force field.

The primary question that this study will address, at least in part, is whether or not it is possible to develop a pairwise additive force field of 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 intermolecular distances (see Equation 7 of Reference 12).

For nearly half a century the Lennard-Jones 12-6 potential has inundated the molecular simulation literature. Only in the past decade has the Mie $n$-6 potential received considerable attention. The popular choice of $n = 12$ (i.e., the Lennard-Jones 12-6 potential) is an historical artifact based primarily on computational reasons that are no longer significant with no theoretical basis. In this study, we propose the development of extended Lennard-Jones (ex-LJ) force fields to improve performance at high pressures.

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

$$u_{\mathrm{nb,exLJ}}(C_m) = \sum_{m=6,8\ldots} C_m r^{-m} \tag{2}$$

where $m$ are integer (typically even) values and $C_m$ are the coefficients for the corresponding $r^{-m}$ 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 $\sigma$ and $\varepsilon$ are the Lennard-Jones size and energy parameters, respectively), while all other $C_m$ values are zero. Also, the $r^{-6}$, $r^{-8}$, and $r^{-10}$ terms can be derived rigorously from attractive London dispersion interactions.[13] By contrast, there is no theoretical basis for the repulsive terms (positive $C_m$, typically $m > 10$).

The ex-LJ potential is more flexible than the two-parameter ($\varepsilon$ and $\sigma$) 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 free energy over the entire fluid region of technological interest? The development of "technical accuracy" ex-LJ force fields allows for prediction of other properties not available from the FEOS, such as transport properties (e.g., self-diffusivity, shear viscosity) and structural/kinetic phenomenon, 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, which combining rules (e.g., Lorentz-Berthelot) should be used for cross interactions? Combining rules are essential for performing simulations of mixtures and compounds with multiple interaction site types.

Due to their theoretical basis, should the $C_6$, $C_8$, and $C_{10}$ coefficients necessarily be negative? Which terms of Equation 2 provide the greatest improvement in the force field? Similarly, how many non-zero terms should be included such that the model is sufficiently flexible but not over-fit?

Although the ex-LJ was proposed over three decades previously,[14] 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 previous hybrid data set studies, the non-bonded parameters were optimized with VLE properties. However, it is unlikely that VLE data alone can provide a unique set of parameters for more than three non-zero $C_m$ terms. By contrast, derivatives of Helmholtz free energy ($A^{\mathrm{r}}_{xy}$) provide high information content regarding the non-bonded potential. Unfortunately, reliable $A^{\mathrm{r}}_{xy}$ values require an accurate FEOS, which is not available *a priori*. For this reason, we propose an iterative hybrid data set approach:

1. Develop FEOS over $P\rho T$ range where reliable experimental data exist
2. Parameterize the ex-LJ force field with FEOS $A^{\mathrm{r}}_{xy}$ over $P\rho T$ region of applicability
3. Iterate:
   (a) Estimate $A^{\mathrm{r}}_{xy}$ for ex-LJ force field at extreme temperatures and pressures
   (b) Refit FEOS to the hybrid data set
   (c) Re-parameterize the ex-LJ force field using updated FEOS $A^{\mathrm{r}}_{xy}$

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 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 3c, although this may necessitate including higher order derivatives ($A^{\mathrm{r}}_{xy}$ for $x+y>2$).

Step 3c 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 $A^{\mathrm{r}}_{xy}$, we propose the use of Multistate Bennett Acceptance Ratio (MBAR) combined with basis functions ($\Phi$). Previous publications demonstrate that MBAR-$\Phi$ reduces the computational cost to estimate ensemble averages by several orders of magnitude compared to direct molecular simulation. Therefore, utilizing MBAR-$\Phi$ in Steps 3c is essential for this algorithm to be computationally tractable. In addition, MBAR-$\Phi$ can be applied in Step 3a to eliminate the need to re-simulate the high temperature and pressure state points for each iteration. Due to the essential role of MBAR-$\Phi$, 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.[15,16] For example, derivatives of the Helmholtz free energy for parameter set $C_m$ are estimated according to

$$A^{\mathrm{r}}_{xy}(C_m) = F_i[\langle U(q_{\mathrm{ref}}, C_m)^i \rangle_{\mathrm{MBAR}}] + F_j[\langle U(q_{\mathrm{ref}}, C_m)^j \rangle_{\mathrm{MBAR}}] \quad (3)$$

where $q_{\text{ref}}$ are configurations sampled using the reference force field(s), $\langle\rangle_{\text{MBAR}}$ are ensemble averages estimated using MBAR (see Equations 9 to 11 of Reference 16), and $F_i$ and $F_j$ are functionals that depend on different powers of the internal energy (see Equations 27 and 30 of Reference 17).

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. Basis functions greatly accelerate the cost of this recalculation step by several orders of magnitude compared to the Gromacs "rerun" function, which is already highly optimized. Due to the linear relationship between the total non-bonded internal energy ($U_{\text{nb,total}}$) and Equation 2, the ex-LJ potential is amenable to basis functions. $U_{\text{nb,total}}$ 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 $N_{\text{sites}}$ 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 $m$. Recomputing $U_{\text{nb,total}}$ for a perturbed set of $C_{m,ij}$ parameters requires simple and *fast* matrix multiplication.

In summary, MBAR-Φ permits rapid estimation of $A^{\text{r}}_{xy}$ for non-simulated extended Lennard-Jones potentials. This renders the iterative hybrid data set approach computationally feasible by removing the need for performing hundreds of molecular simulations.

Four deliverables are expected from this project. First, a "technical accuracy" FEOS for each molecule studied. Second, an extended Lennard-Jones force field of "technical accuracy." Third, an increased theoretical understanding of non-bonded potentials. Fourth, an infrastructure for rapid force field parameterization using Helmholtz free energy derivatives.

## References


[1] G. Rutkai, M. Thol, R. Lustig, R. Span, and J. Vrabec. Communication: Fundamental equation of state correlation with hybrid data sets. *The Journal of Chemical Physics*, 139(4):041102, 2013.

[2] M. Thol, F.H. Dubberke, G. Rutkai, T. Windmann, A. Köster, R. Span, and J. Vrabec. Fundamental equation of state correlation for hexamethyldisiloxane based on experimental and molecular simulation data. *Fluid Phase Equilibria*, 418:133 – 151, 2016. Special Issue covering the Nineteenth Symposium on Thermophysical Properties.

[3] Monika Thol, Gábor Rutkai, Andreas Köster, Frithjof H. Dubberke, Thorsten Windmann, Roland Span, and Jadran Vrabec. Thermodynamic properties of octamethylcyclotetrasiloxane. *Journal of Chemical & Engineering Data*, 61(7):2580–2595, 2016.

[4] Monika Thol, Gábor Rutkai, Andreas Köster, Svetlana Miroshnichenko, Wolfgang Wagner, Jadran Vrabec, and Roland Span. Equation of state for 1,2-dichloroethane based on a hybrid data set. *Molecular Physics*, 115(9-12):1166–1185, 2017.

[5] Monika Thol, Gábor Rutkai, Andreas Köster, Mirco Kortmann, Roland Span, and Jadran Vrabec. Fundamental equation of state for ethylene oxide based on a hybrid dataset. *Chemical Engineering Science*, 121:87 – 99, 2015. 2013 Danckwerts Special Issue on Molecular Modelling in Chemical Engineering.

[6] Monika Thol, Gábor Rutkai, Andreas Köster, Rolf Lustig, Roland Span, and Jadran Vrabec. Equation of state for the Lennard-Jones fluid. *Journal of Physical and Chemical Reference Data*, 45(2):023101, 2016.

[7] Monika Thol, Gábor Rutkai, Roland Span, Jadran Vrabec, and Rolf Lustig. Equation of state for the Lennard-Jones truncated and shifted model fluid. *International Journal of Thermophysics*, 36(1):25–43, Jan 2015.

[8] Gábor Rutkai, Monika Thol, Roland Span, and Jadran Vrabec. How well does the Lennard-Jones potential represent the thermodynamic properties of noble gases? *Molecular Physics*, 115(9-12):1104–1121, 2017.

[9] Rolf Lustig, Gábor Rutkai, and Jadran Vrabec. Thermodynamic correlation of molecular simulation data. *Molecular Physics*, 113(9-10):910–931, 2015.

[10] Gábor Rutkai and Jadran Vrabec. Empirical fundamental equation of state for phosgene based on molecular simulation data. *Journal of Chemical & Engineering Data*, 60(10):2895–2905, 2015.

[11] Gábor Rutkai, Andreas Köster, Gabriela Guevara-Carrion, Tatjana Janzen, Michael Schappals, Colin W. Glass, Martin Bernreuther, Amer Wafai, Simon Stephan, Maximilian Kohns, Steffen Reiser, Stephan Deublein, Martin Horsch, Hans Hasse, and Jadran Vrabec. ms2: A molecular simulation tool for thermodynamic properties, release 3.0. *Computer Physics Communications*, 221:343 – 351, 2017.

[12] Michał Przybytek, Wojciech Cencek, Bogumił Jeziorski, and Krzysztof Szalewicz. Pair potential with submillikelvin uncertainties and nonadiabatic treatment of the halo state of the helium dimer. *Phys. Rev. Lett.*, 119:123401, Sep 2017.

[13] Anthony Stone. The theory of intermolecular forces, 2nd edition. 54, 07 2013.

[14] Ferenc Kalos and Arthur E. Grosser. Intermolecular potentials from differential cross sections: Ar + Ar. *Canadian Journal of Chemistry*, 50(6):892–896, 1972.

[15] Michael R. Shirts and John D. Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of Chemical Physics*, 129:124105, 2008.

[16] 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. PMID: 29727563.

[17] Rolf Lustig. Statistical analogues for fundamental equation of state derivatives. *Molecular Physics*, 110(24):3041–3052, 2012.