Density calculation for organics using pcff+, in HT mode

In 1964, in his *Lectures on Physics*, Richard Feynman reflected *"If we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms."*

Studying matter at the atomic scale can indeed provide invaluable information on how matter behaves and why it behaves the way it does. Forcefield simulations are widely used nowadays on systems containing hundreds to millions of atoms and for times that extend from a few picoseconds to microseconds.

Density is a fundamental macroscopic property that can be calculated by forcefield simulations and which needs to be in excellent agreement with experimental data for any other property prediction to be meaningful. Other macroscopic properties of interest for fluids that may be calculated are saturation pressure, vaporization enthalpy, normal boiling point, critical point, solubility, diffusivity, viscosity, and thermal conductivity.

In this work, we use molecular dynamics (MD) simulations, with the PCFF+ forcefield [1], to calculate saturated liquid density at several temperatures between the melting and the critical point of pure organic compounds. The calculations are performed in High-Throughput (HT) mode to facilitate and expedite setting up simulations as well as collecting and post-processing simulation output. A well-established and straightforward Group-Contribution QSPR method [2] is used to select a set of temperatures for performing simulations for each compound.

1 Forcefields

Forcefield simulations rely—as the name implies—on the use of forcefields to describe the interatomic (or interparticle) interactions. In the case of molecular systems, these interactions are both intra- and intermolecular.

For the past four decades, numerous forcefields have been developed and proposed in the literature. Based on the forcefield development approach and techniques, the terms included, the energy functional forms used, the assumptions and approximations made, they can be classified as, e.g., Class I [7] [8] [9], Class II [11] [3] [4], Class III, Embedded Atom (EAM) [5], ReaxFF [6], and Machine Learning (ML) [13].

[1] Rigby, D. "The PCFF+ forcefield for condensed matter simulation: Overview and validation", in preparation.

[2] Joback, K. G.; Reid, R. C. "Estimation of pure-component properties from Group-Contributions", *Chem. Eng. Comm.* **57**, 233–243, (1987) https://doi.org/10.1080/00986448708960487

[7] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell, A. D., Jr., "CHARMM General Force Field: A Force Field for Drug-like Molecules Compatible with the CHARMM All-Atom Additive Biological Force Fields", *J Comput Chem*, **31**, 671–690, (2010) https://doi.org/10.1002/jcc.21367

[8] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., "A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules", *J. Am. Chem. Soc.* **117**, 5179–5197, (1995) https://doi.org/10.1021/ja00124a002

[9] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J., "Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids", *J. Am. Chem. Soc.* **118**, 11225–11236, (1996) https://doi.org/10.1021/ja9621760

[11] Maple, J. R.; Hwang, M.-J.; Stockfisch, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T., "Derivation of Class II Force Fields. I. Methodology and Quantum Force Field for the Alkyl Functional Group and Alkane Molecules", *Journal of Computational Chemistry* **15**, 162–182, (1994) https://doi.org/10.1002/jcc.540150207

[3] Sun, H.; Mumby, S. J.; Maple, J. R.; Hagler, A. T., "An Ab Initio CFF93 All-Atom Force Field for Polycarbonates", *J. Am. Chem. Soc.* **116**, 2978–2987, (1994) https://doi.org/10.1021/ja00086a030

[4] Sun, H. "COMPASS An Ab Initio Force-Field Optimized for Condensed-Phase Applications Overview with Details on Alkane and Benzene Compounds". *J. Phys. Chem. B*, **102**, 7338–7364, (1998) https://doi.org/10.1021/jp980939v

[5] Daw, M. S.; Baskes, M. I., "Embedded-Atom Method: Derivation and Application to Impurities, Surfaces, and Other Defects in Metals", *Phys. Rev. B*, **29**, 6443–6453, (1984) https://doi.org/10.1103/PhysRevB.29.6443

[6] Senftle, T. P.; Hong, S.; Islam, M. M.; Kylasa, S. B.; Zheng, Y.; Shin, Y. K.; Junkermeier, C.; Engel-Herbert, R.; Janik, M. J.; Aktulga, H. M.; Verstraelen, T.; Grama, A.; van Duin, A. C. T., "The ReaxFF Reactive Force-Field: Development, Applications and Future Directions", *npj Computational Materials* **2**, 15011, (2016) https://doi.org/10.1038/npjcompumats.2015.11

[13] Unke, O. T.; Chmiela, S.; Sauceda, H. E.; Gastegger, M.; Poltavsky, I.; Schütt, K. T.; Tkatchenko, A.; Müller, K.-R., "Machine Learning Force Fields", *Chem. Rev.* 2021. https://doi.org/10.1021/acs.chemrev.0c01111

Important aspects of any forcefield that define their applicability and their ability to be used for property prediction are:

- Quality of property prediction
- Chemical compounds' breadth of coverage
- Transferability of parameters
- Range of properties that can be studied

2 PCFF+ description

The PCFF+ forcefield [Page 1, 1] is an extension of PCFF [Page 1, 3], a Class II, All-Atom (AA) forcefield that has its origin in the CFF series of forcefields [Page 1, 11].

$$E = \sum E^b + \sum E^a + \sum E^0 + \sum E^t + \sum E^{bb} + \sum E^{ab} + \sum E^{aa} + \sum E^{at} + \sum E^{bt} + \sum E^{elec} + \sum E^{VDW} \tag{2.1}$$

where

$$E^b = \sum_{i=2}^{4} k_i^b (b - b_0)^i \tag{2.2}$$

$$E^a = \sum_{i=2}^{4} k_i^a (\theta - \theta_0)^i \tag{2.3}$$

$$E^t = \sum_{i=1}^{4} k_i^t (1 - cosi\phi) \tag{2.4}$$

$$E^0 = k^0 (\chi - \chi_0)^2 \tag{2.5}$$

$$\{E^{bb}, E^{aa}, E^{ab}\} = k^c (s - s_0)(s' - s'_0) \tag{2.6}$$

$$\{E^{bt}\} = (b - b_0) \sum_{i=1}^{3} k_i^c (1 - cosi\phi)\} \tag{2.7}$$

$$\{E_{at} = (\theta - \theta_0) \sum_{i=1}^{3} k_i^c (1 - cosi\phi)\} \tag{2.8}$$

$$E^{el} = \sum_{ij} \frac{q_i q_j}{r_{ij}} \tag{2.9}$$

$$E^{VDW} = \sum_{ij} \epsilon_{ij} \left[2 \left(\frac{r^*_{ij}}{r_{ij}} \right)^9 - 3 \left(\frac{r^*_{ij}}{r_{ij}} \right)^6 \right] \tag{2.10}$$

are the bond E^b, angle E^a, torsion E^t, out-of-plane angles E^0, bond-bond E^{bb} / angle-angle E^{aa} / bond-angle E^{ba}, bond-torsion E^{bt}, electrostatic E^{el}, and van der Waals terms E^{VDW}, respectively. The nonbond terms (van der Waals and electrostatic) are pairwise interactions between atoms that are separated by three or more bonds (intramolecular interactions) that belong to different molecules (intermolecular interactions).

Amongst the different energy contributions, the nonbond interactions are often the most difficult to parameterize, as they cannot be directly obtained from ab initio calculations. PCFF+ extensions and improvements on the original PCFF forcefield focus on precisely this point, i.e. the optimization of existing nonbond parameters as well as the introduction and parameterization of new atom types to achieve the highest possible coverage and accuracy for organic (and some inorganic) species, while maintaining forcefield transferability and a manageable number of distinct atom types.

PCFF+ provides coverage for all major groups of organic species, shown in figure Figure 2.1. As is shown, there is availability of all required forcefield parameters for 95% of the species included in DIPPR [10], one of the largest molecular databases. The vast majority of the database consists of organic compounds, but there are also many inorganic compounds.

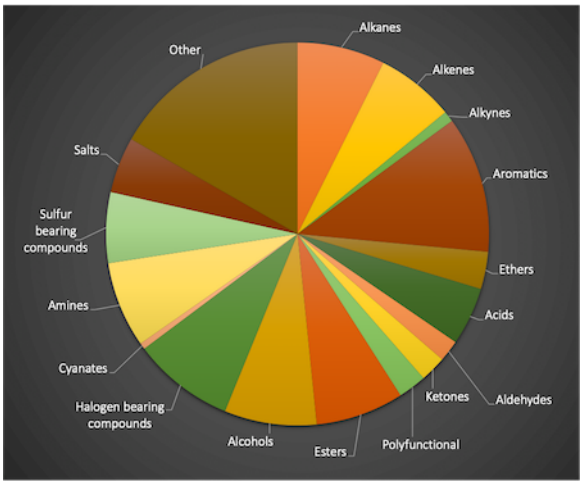


Figure 2.1: Chemical families of compounds tested for forcefield coverage with PCFF+. A total of 2,150 compounds are included in the coverage analysis.

[10] Wilding, W. V.; Rowley, R. L.; Oscarson, J. L., "DIPPR® Project 801 Evaluated Process Design Data", *Fluid Phase Equilibria* **150–151**, 413–420, (1998) https://doi.org/10.1016/S0378-3812(98)00341-0

3 Molecular Simulations

A significant benefit of the use of forcefield simulations for property prediction is the fact that the amount of human time needed to set up the simulations can be minimized with the use of robust High-Throughput (HT) computation protocols and reliable compute engines.

Here we use a robust MedeA HT computational protocol to calculate saturated liquid density at five (5) temperatures for twenty-eight (28) compounds. The only input provided by the user is a list of compounds, together with their SMILES codes and the number of temperatures at which to run the simulations. Everything is processed in a single job, a single workflow, which sets up different tasks for creating the appropriate input (including the creation of amorphous liquid-like configurations of each compound and the choice of temperatures), submitting the computations, retrieving the output once a task is finished, post-processing the output and printing the results in the desired format. Depending on the availability of computing resources, the different tasks can run in a serial or parallel way, locally or remotely.

The schematic workflow of the entire job is shown in Figure 3.1.

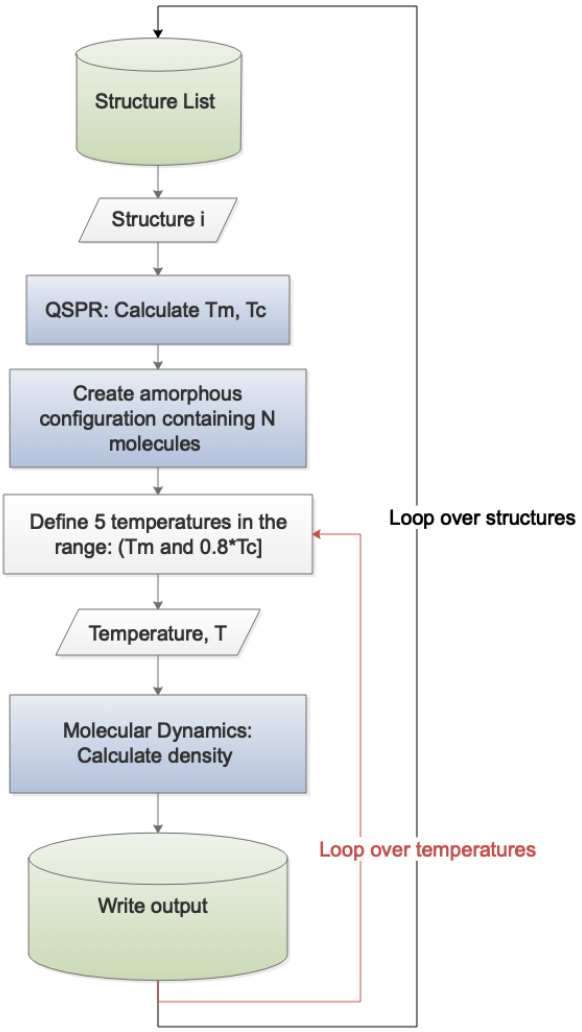


Figure 3.1: Workflow of a single job for calculating saturated liquid densities at five temperatures and 1 atm, for 28 organic compounds.

4 Computational Details

The first step is the definition of a list of compounds containing the name and SMILES string.

Then, for each compound in this list, the Joback Group-Contribution method is used to estimate the melting temperature (T_m) and the compound's critical temperature (T_c). Five temperatures are then selected, in the range $(T_m, 0.8 * T_c]$.

An amorphous configuration containing several molecules is created for each compound often at a relatively low density of 0.3 g/ml, though higher densities may optionally be used. The number of molecules is set so that at least 1,200 atoms are present in the system.

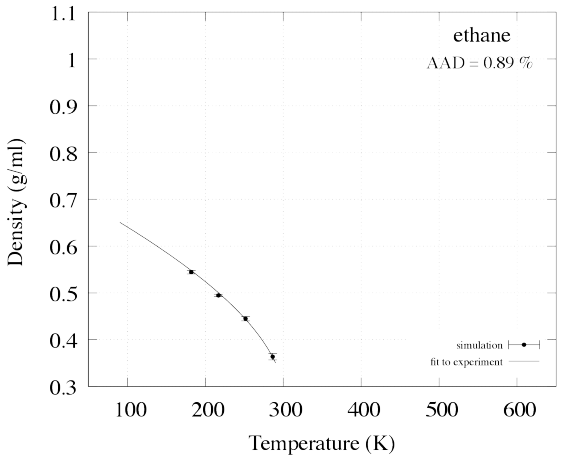
For each temperature, an MD simulation is started, including a short NVT run (100 ps) followed by a longer NPT run (1 ns). Density is calculated from the simulation output for the time interval for which convergence has been reached after a convergence analysis is performed.

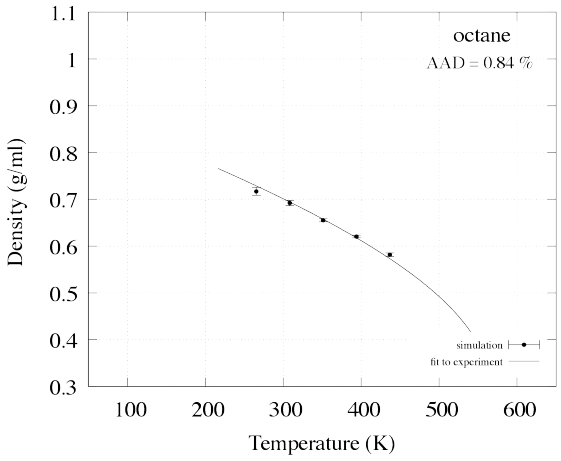
Results are collected, and the appropriate tables are printed, together with a summary of the computation.

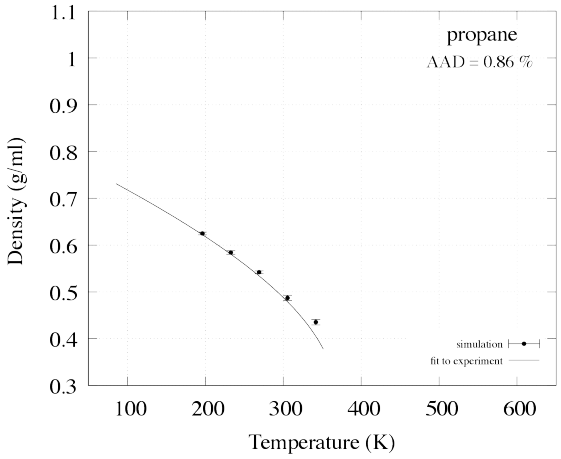
Human intervention is not required after submitting the workflow (as a single job) to the compute resources. Moreover, the number of compounds may be increased, as appropriate, without any change to the protocol or the human time needed for setting up the workflow.

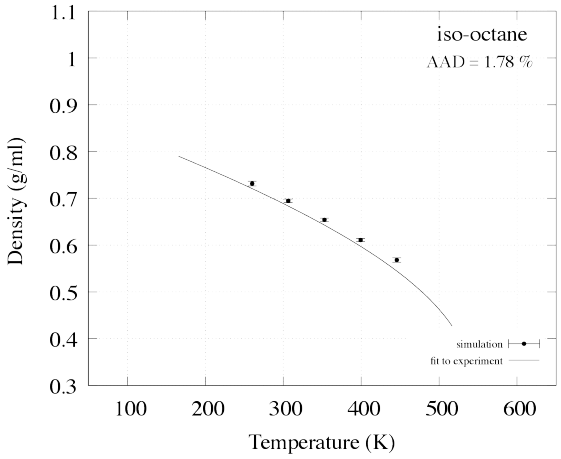
5 Results and Discussion

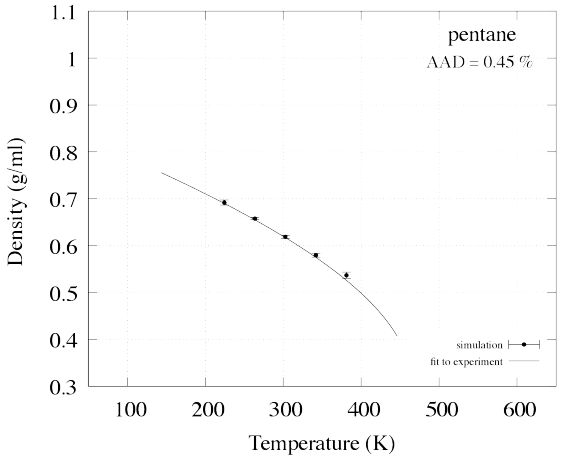
PCFF+ based density calculations over wide ranges of temperature are presented in the figures below and compared with curves fitted to curated experimental data as developed within the DIPPR 801 database project [Page 2, 10] .

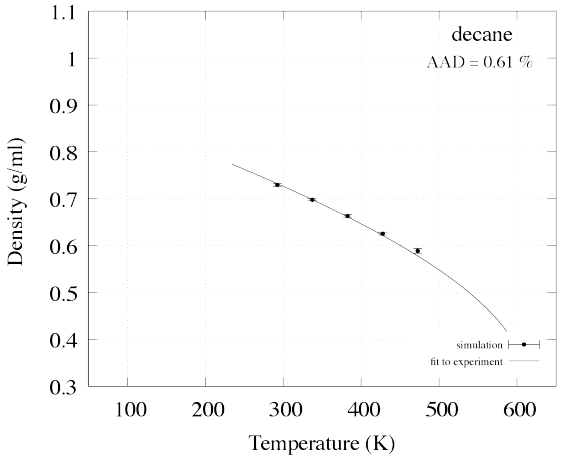


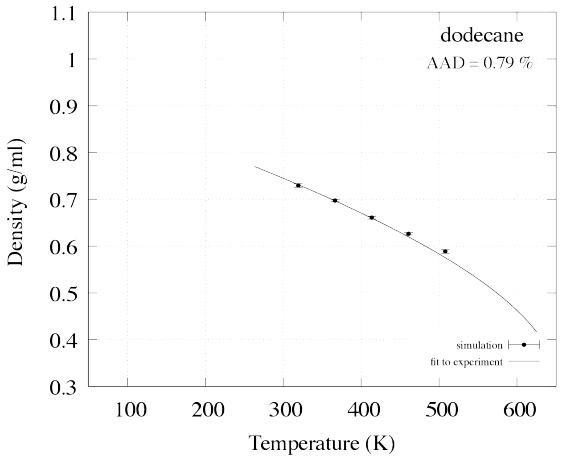


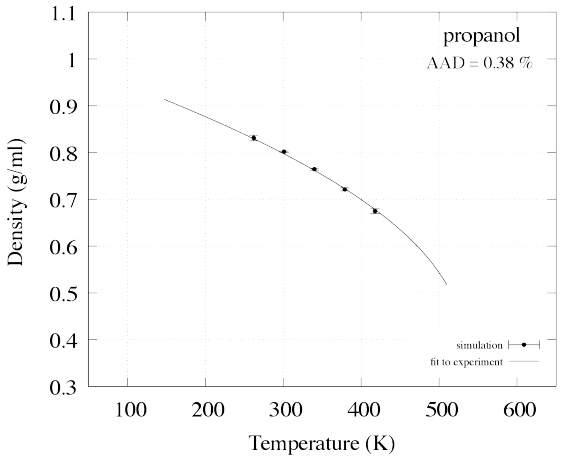


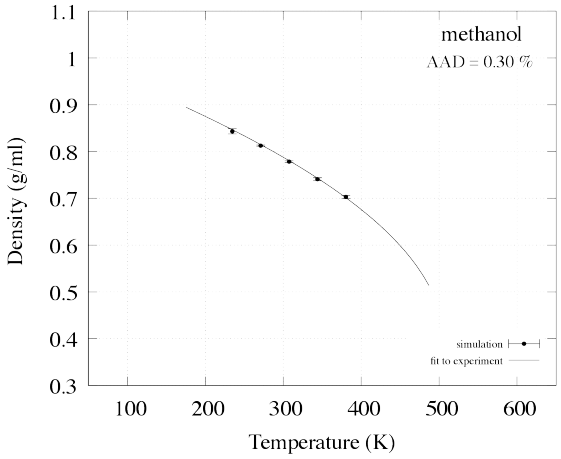


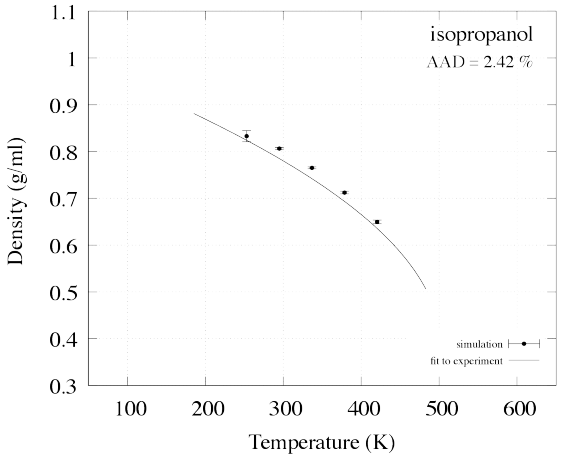


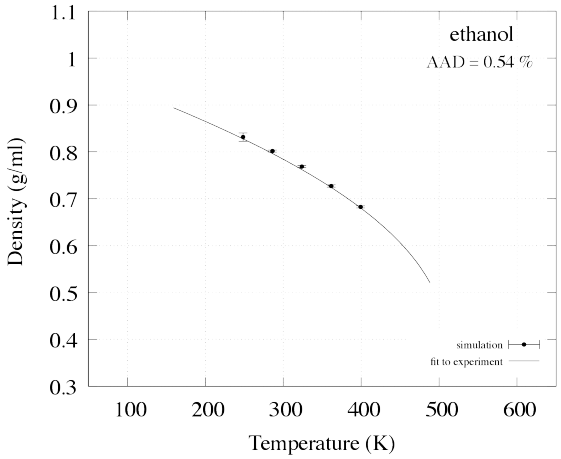


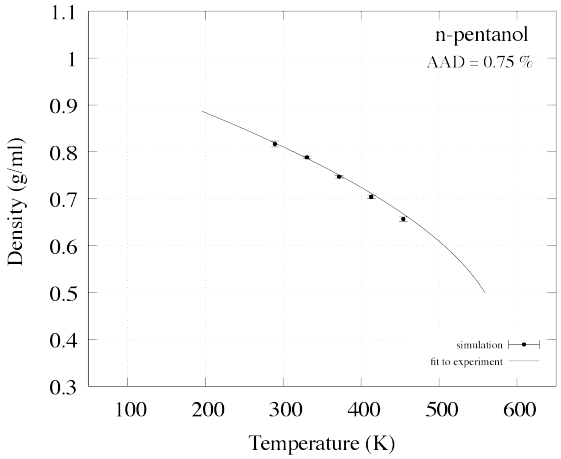


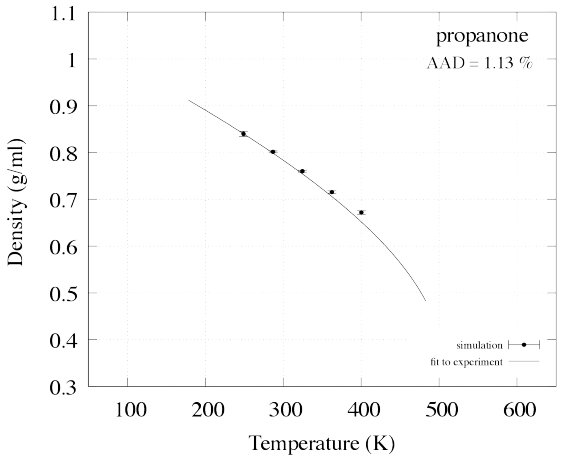


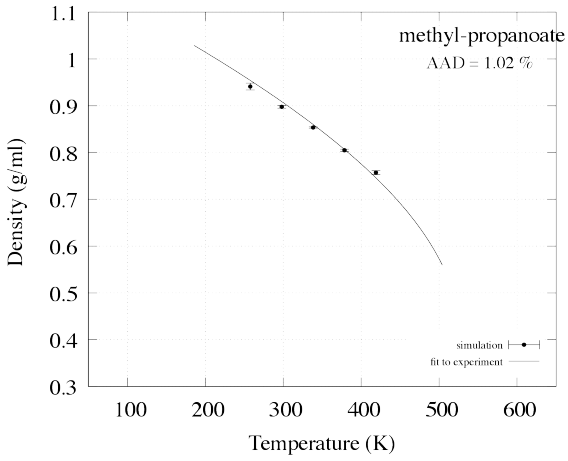


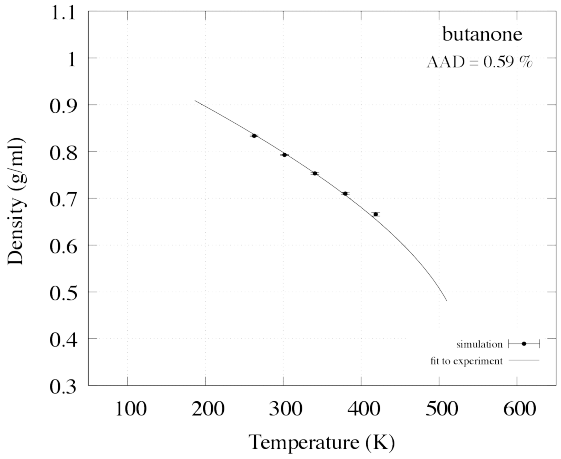


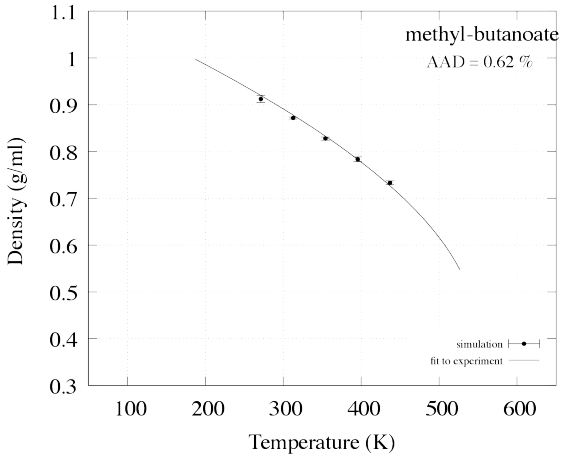


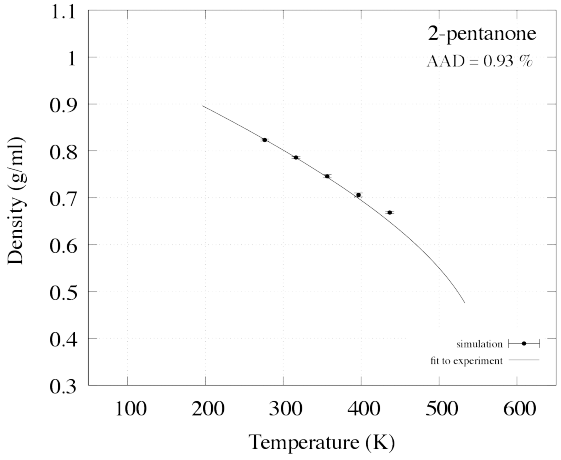


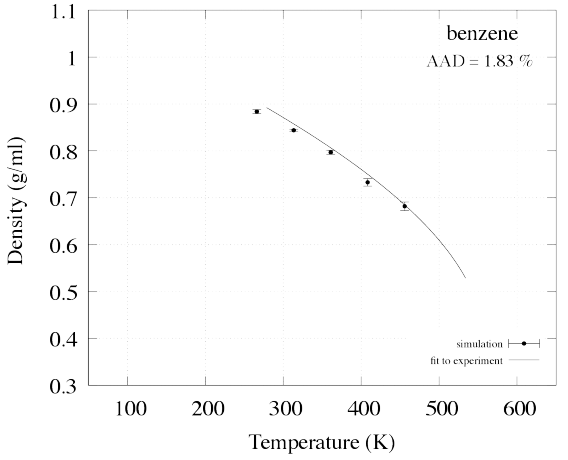


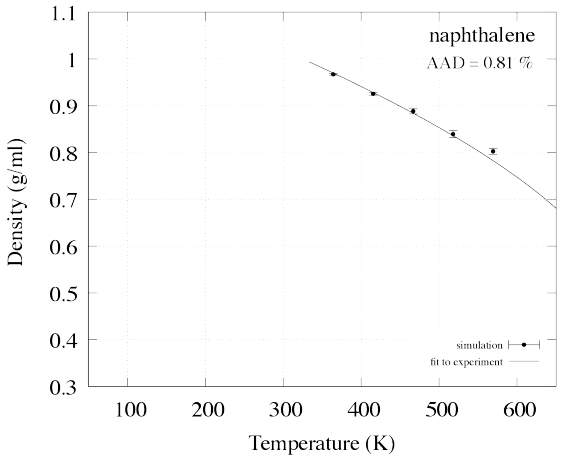


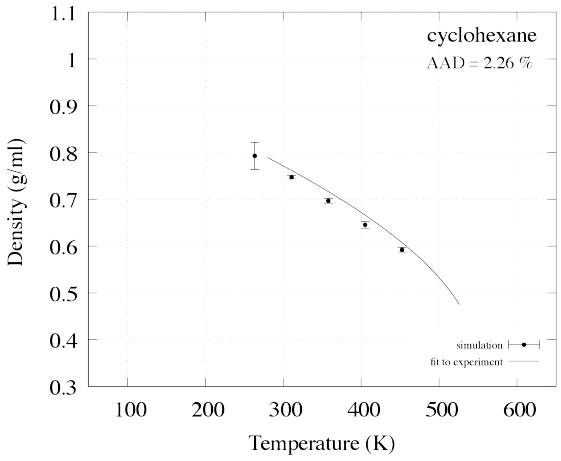


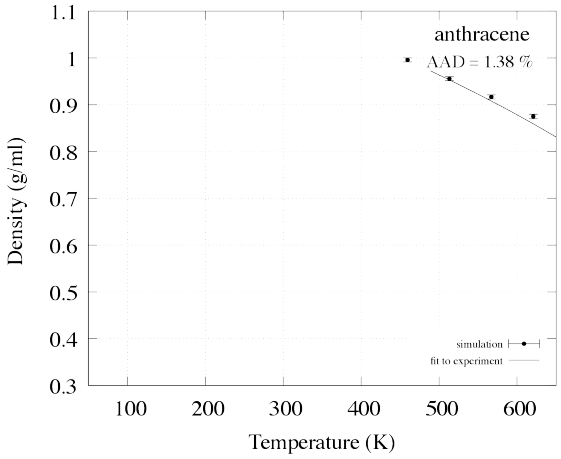


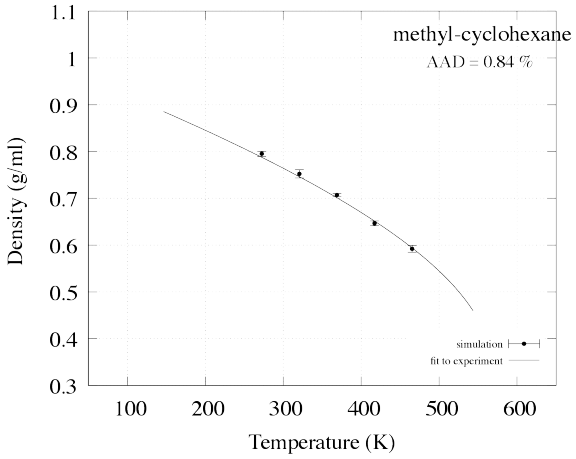


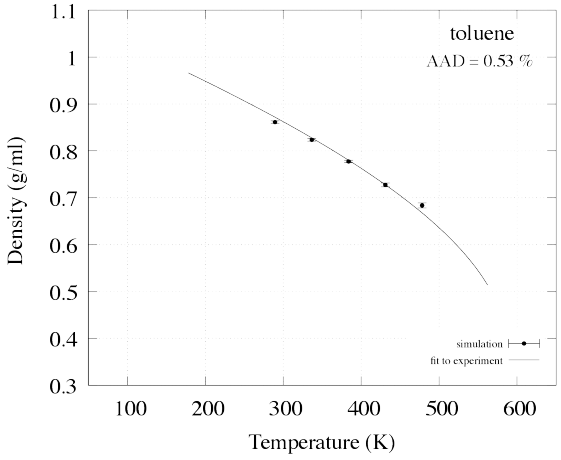


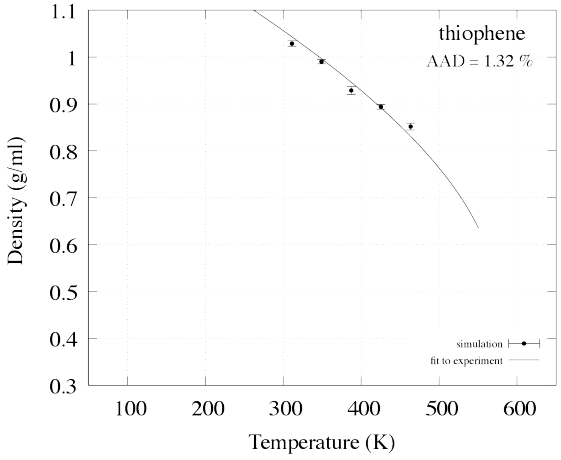


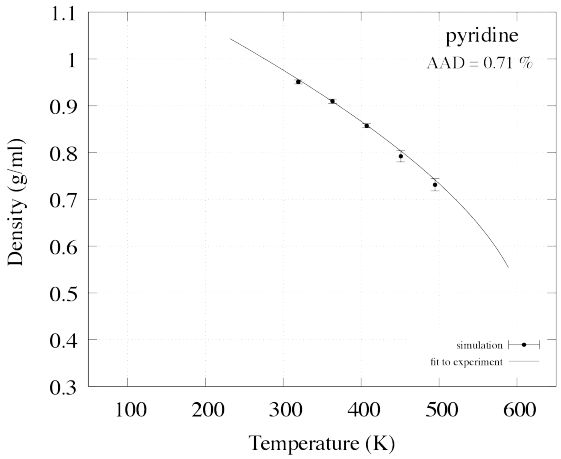


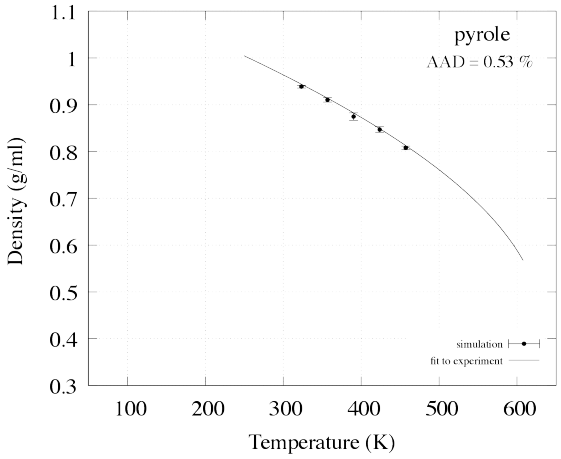


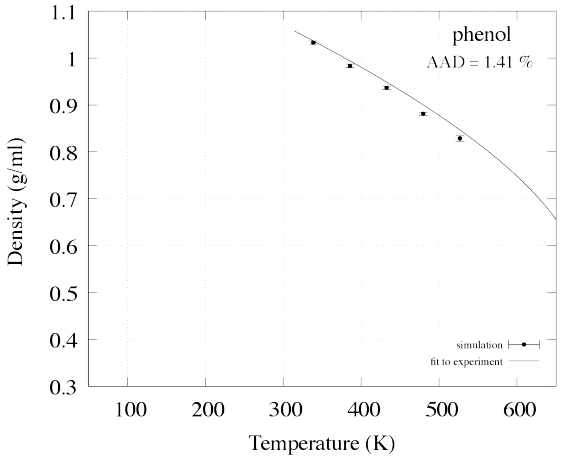












Overall, agreement with the correlated experimental data is excellent, generally exhibiting average absolute deviations better than the parameterization target of 1% between the melting temperature and normal boiling point imposed for all new parameter development in PCFF+.