

A coarse-grained parametrisation of linear alkanes.

Otello M Roscioni¹, Matteo Ricci¹, Stephen Farr²

¹MaterialX LTD, Bristol, United Kingdom.

²EPCC - The University of Edinburgh, United Kingdom.

Abstract

The main goal of this project was to develop a computational workflow for calculating the vaporisation enthalpies of pure compounds. Specifically, we investigated the family of linear alkanes from C2 to C16 using the MOLC model, a coarse-grained force field developed by MaterialX and implemented in the open-source molecular dynamics code LAMMPS. The first task of this project was creating new software for calculating the intermolecular energy, while the second was to devise a robust parametrisation strategy yielding accurate thermochemical properties. The calculations showed that the Gay-Berne potential energy well ϵ_0 needs to be scaled to compensate for the superposition of ellipsoidal beads in the MOLC model of linear alkanes. This result represented an important technical advancement in developing the MOLC model and laid the basis for creating a thermodynamically and structurally accurate coarse-grained force field for lipids and surfactants.

Introduction

MaterialX spun out from the Horizon 2020 project EXTMOS, in which the founders developed original software for molecular modelling. It was incorporated in February 2020 and currently operates with a consultancy business model. MaterialX's main focus is maintaining a coarse-grained (CG) model called MOLC¹, and a suite of open-source software for executing multiscale simulations of soft materials such as organic semiconductors, polymers, and solvents. A consistent part of this effort consists of creating a portfolio of use cases where MaterialX's technology is applied to study various materials. With the SHAPE project, MaterialX aimed to implement a new protocol for the parametrisation of the MOLC force field and to test it for the modelling of linear alkanes. This choice was motivated by the fact that alkanes are ubiquitous functional units used to increase the solubility of polymers and are the main structural element in fats and lipids. MaterialX benefitted from the HPC facilities and the technical support made available through the SHAPE project. The project was successfully completed on September 2022, and the most important outcomes are the following:

- Refactoring of the old MOLC user package to a new pair style, with increased computational efficiency.
- Porting the MOLC user package to the latest LAMMPS development version available during the project timeframe (2 June 2022).
- Creating a LAMMPS module to compute the intermolecular energy, with appropriate corrections for long-range Coulombic potential.
- Benchmarking of new software against reference simulations and experimental data.
- Parametrisation and characterisation of a new coarse-grained force field for linear alkanes based on the MOLC model.

- Application of the improved parametrisation strategy for modelling poly lactic-co-glycolic acid (PLGA) based on the MOLC model.

The outcomes of this project that are potentially relevant for the LAMMPS community have also been presented on the topic of [intermolecular potential energy](#) on the [MatSci](#) forum.

Theoretical background

The calculation of the intermolecular energy in the case of bonded systems with long-range solvers requires computing the pairwise interaction energy for all particles in the system and subtracting the intramolecular contribution to the total pairwise energy. For truncated potentials, this operation is trivial. Potentials that decay slowly with the distance, such as the Coulombic potential, are usually treated with long-range solvers, making the intermolecular energy calculation more complex. For instance, in an Ewald sum^{2,3}, the total Coulombic potential U_C is divided into a real space portion U_C^{real} and reciprocal k -space portion U_C^{recip} as:

$$U_C = U_C^{real} + U_C^{recip} - U_C^{self} \quad (1)$$

where U_C^{self} is a correction term to remove the interaction of a charge with itself in the replicated cells⁴. The real-space portion of the Coulombic energy is calculated for $r \leq r_{cut}$ and includes an additional damping factor to compensate for the use of long-range solvers:

$$U_C^{real} = \frac{1}{4\pi\epsilon_0} \sum_{i < j} \frac{q_i q_j}{r_{ij}} \text{erfc}(\alpha r_{ij}) \quad (2)$$

where $\text{erfc}(x)$ is the complementary error function, and α is the Ewald parameter that allows the balancing of calculations between the direct and reciprocal space.

The intermolecular energy is the pairwise energy between atoms in different molecules. In LAMMPS, an existing `compute group/group` command computes the intermolecular energy between atoms, with the additional constraint that they must be in different molecules. This command is equivalent to computing the pairwise energy by excluding all intramolecular contributions as if the following command was issued:

```
neigh_modify exclude molecule/intra all
```

The net result is that the intramolecular energy term U_C^{real} is set to zero, while the reciprocal k -space portion U_C^{recip} is unaffected by exclusions. The problem of applying this procedure to potentials with a long-range component is that the intramolecular energy includes a correcting factor, while the intramolecular energy of an isolated molecule should be independent of whether or not a long-range solver is used. Indeed, the unscaled Coulombic potential is subtracted from Eq. (2) for *bonded pairs* via the `special_bonds` command, yielding:

$$U_C^{real} = \frac{1}{4\pi\epsilon_0} \sum_{i < j} \frac{q_i q_j}{r_{ij}} (\text{erfc}(\alpha r_{ij}) - 1 + w_c) \quad (3)$$

where w_c is a weight factor that goes from zero (no interaction) to one (fully interacting).

Therefore, the correct way of excluding the intermolecular energy contribution for systems with a long-range Coulombic solver is to use Equation (3) with $w_c = 0$ for every pair

belonging to the same molecule, and $w_c = 1$ otherwise. On top of that, the MOLC model requires looping over every charge belonging to the same bead, which is not required in atomistic models.

In this project, we forked the code of the `compute group/group` command into a new `compute intra` command, which acts on a single group of atoms and computes the real part of the Coulomb summation with a weight factor equal to 0 for every atom pair belonging to the same molecule, as in Eq. (3). The `compute intra` command has been developed and tested first for bonded atomic systems, where every particle carries a charge, and then for CG systems based on the MOLC model, where each particle can carry more than one charge. For the CG model, the `compute intra/molc` command has an extra loop to account for the interaction between point charges inside the same bead, but it is otherwise identical to the `compute` for atomic systems.

MOLC model

The MOLC CG model¹ is based on replacing bonded molecular fragments with biaxial Gay-Berne ellipsoids decorated with massless point charges. In large molecules, ellipsoids are connected with directional bonds, and they overlap in space: as a result, the CG molecules maintain some internal degrees of freedom, such as rotations around selected bonds, and reproduce the molecular excluded volume accurately. This unique characteristic of the MOLC model allows to back-map CG molecules back to their former atomic structure without loss of spatial information. The molecular packing obtained with the MOLC model yields accurate mass density and other condensed-phase properties for both amorphous solids and liquids. However, the trade-off of this approach is that pairwise interactions between bonded pairs need to be zeroed. This leads to a tighter packing of Gay-Berne ellipsoids in space, resulting in a systematic overestimate of the intermolecular energy. The overestimation of intermolecular energy increases with the degree of superposition and the number of bonds in a given molecule. This effect is graphically illustrated by considering the series of linear alkanes shown in Figure 1.

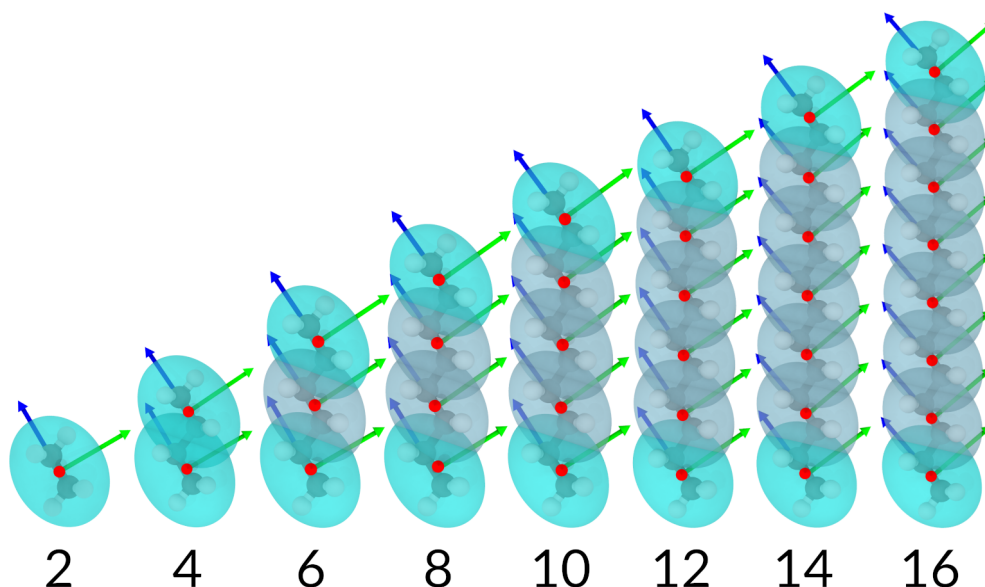


Figure 1: Coarse-grained models of linear alkanes (blue ellipsoids) superimposed with their atomic structure (ball-and-stick) and axes of symmetry showing their orientation. The number indicates the alkane's length in terms of carbon atoms.

In order to compensate for this error, the proposed strategy is to scale the energy well depth ε_0 while keeping fixed other parameters, such as the size of the ellipsoids and the electrostatic potential on each bead. The experimental enthalpy of vaporisation is used as a reference to calibrate the energy well. Assuming that the intramolecular and kinetic energy of molecules in the liquid and gas phases is the same⁵, the enthalpy of vaporisation can be written as a function of the intermolecular energy of the liquid phase:

$$\Delta H_{vap}(T) = -E_{inter}^{liquid}(T) + RT \quad (4)$$

This strategy was used to optimise the force field parameters for the family of linear alkanes with an even number of carbon atoms going from ethane (C2) to hexadecane (C16), for which a rich collection of thermodynamical properties are available in the NIST ThermoData Engine (TDE103a v10.4.2)⁶.

The parametrisation strategy can be made more general by directly comparing the intermolecular energy of a CG sample with a reference value computed from a higher-resolution model such as an atomistic force field or *ab-initio*. This latter approach can be applied to materials for which experimental data are hard to determine, such as the cohesive energy of polymers.

This report primarily deals with the parametrisation and characterisation of a CG model of linear alkanes. A brief description of a polymer's parametrisation is also reported.

Linear alkanes

Several CG models of linear alkanes have been reported in the literature⁷⁻⁹, where many carbon atoms are grouped into spherical beads connected with bond and angle potentials. The CG model presented in this work uses uniaxial ellipsoids to represent two carbon units. The pairwise interactions in the MOLC model are a combination of a short-range Gay-Berne potential plus a long-range Coulombic potential based on massless point charges. The Gay-Berne parameters have been modelled on the ethane molecule, with the 3 diameters σ_i and energy well ε_0 optimised to reproduce the experimental enthalpy of vaporisation and density at 129 K. These parameters have been used in longer alkanes to describe CH_3CH_2 and CH_2CH_2 units. Due to this choice, only linear alkane with an even number of carbon atoms is considered in this study.

6 point charges have been used to describe the molecular electrostatic potential of ethane. For longer alkanes, 4 point charges were used for the terminal CH_3CH_2 group, while central CH_2CH_2 units had no charges attached. The position of the point charges relative to the parent ellipsoid was kept constant, while the magnitude of the charges was fitted on the electrostatic reference potential, as computed with the semiempirical molecular-orbital program MOPAC¹⁰, part of the Advanced Topology Builder (ATB) repository¹¹, version 3.0. The same set of reduced charges was used for alkanes longer than 12 carbon units since no significant electrostatic potential variation was observed.

The bonded potential has been modelled on the torsional potential between carbon atoms in the butane molecule, shown in Fig. 2. The potential energy of the butane molecule was used to map every configuration to the scalar products between the axes of inertia of the molecular fragments. The potential energy was computed using the GROMOS-ATB force field¹¹, scaling the energy relative to the minimum and scaling each component by 1/16, hence assuming that the energy is equally distributed on the CG degrees of freedom. The CG force field with the default set of parameters yields a good mass density for all linear alkanes. However, the intermolecular energy is increasingly overestimated for longer

chains. To compensate for the superposition of Gay-Berne ellipsoids forming the CG molecules, the experimental enthalpy of vaporisation (from NIST TDE⁶) was used to scale the energy well ε_0 .

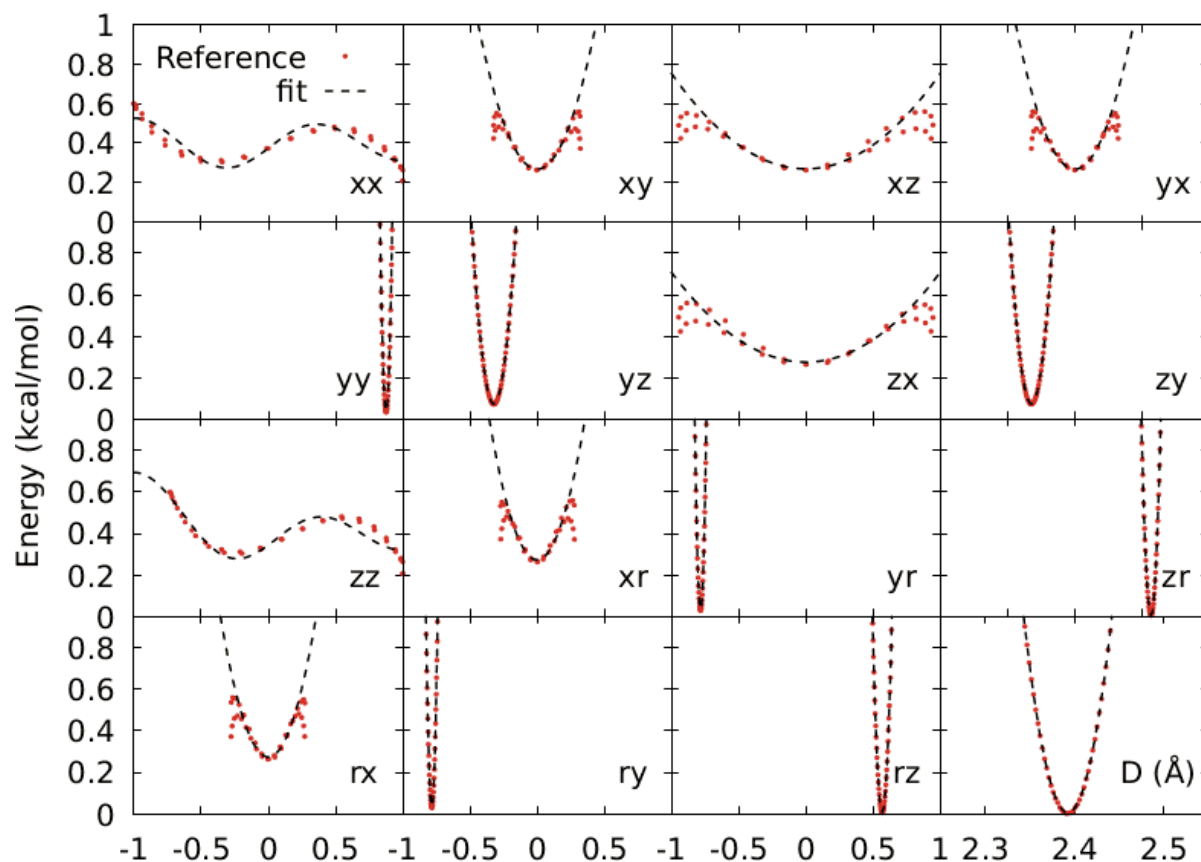


Figure 2: Potential energy of the two beads representing *n*-butane (C_4H_{10}) during the rotation around the C1-C2-C3-C4 dihedral angle, broken down into the 15 scalar products between axes and bond vector, plus the distance between beads.

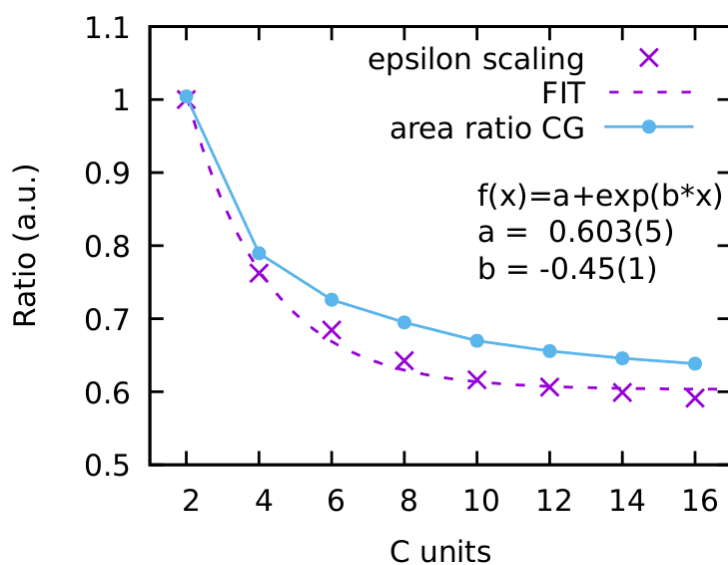


Figure 3: Scaling the energy well ε_0 as a function of the alkane's length. The CG area ratio is computed between the surface area of the CG molecule and the sum of the areas of individual ellipsoids.

The enthalpy was computed using Eq. (4) at a temperature where the liquid phase is stable at 1 atm pressure. The ratio between the optimised energy wells ε_0 and the one optimised for ethane is shown in Fig. 3. The scaling factor shows an exponential trend that converges to a limit value of 0.6.

The plot also shows a good correlation between the scaling factor and the ratio between the surface area of CG molecules and the sum of the areas of the individual ellipsoids describing the molecule. The CG area ratio could be used to scale the energy well depth in the absence of reference data. Another way is to produce synthetic data, e.g. by back-mapping a CG sample of a pristine compound into an atomistic one, and then directly use the intermolecular energy to compute the scaling factor.

The optimised CG force field has been used to compute the following physical observables in a temperature range covering the area of existence of the liquid phase: density, enthalpy (or heat) of vaporisation, viscosity, surface tension, and self-diffusion coefficient.

Results

The CG samples of linear alkanes were obtained by placing 1000 randomly-oriented molecules on a cubic grid with long enough spacing to avoid bad contacts, with 3D periodic boundary conditions applied. The samples were compressed to the experimental mass density at room temperature and further equilibrated in isobaric-isothermal conditions (NPT ensemble) until reaching thermal equilibrium. The temperature was controlled via a Langevin thermostat with a damping parameter of 20 ps and the pressure with a Nosé-Hoover barostat with a stress-damping parameter of 200 ps. The equations of motion were integrated with a timestep of 20 fs for a production time of 10 ns. 10 simulations were performed with the sample being heated or cooled, and the results averaged together to minimise hysteresis effects. The density was computed directly from the thermodynamic output of LAMMPS. The vaporisation enthalpy was computed post-processing from the sample's actual temperature and its intermolecular energy. The viscosity was calculated using the Green-Kubo formula, based on the ensemble average of the autocorrelation function of the stress/pressure tensor computed in the canonical ensemble (NVT). The surface tension was calculated using the Irving-Kirkwood relation, based on the normal and tangential pressure difference due to the surface's formation in a free-standing solvent film. The self-diffusion coefficient was calculated using Einstein's relation from the slope of the mean-squared displacement over time. The physical observables were computed over a range of temperatures covering the liquid phase of each alkane, with a resolution of 10 K. Each temperature step was obtained by changing the temperature at a rate of 2.5 K/ns, followed by 10 ns of equilibration and 10 ns of production time.

The results in Fig. 4 show a good agreement with experimental data over a wide range of temperatures, with the density, heat of vaporisation, and self-diffusion coefficient reaching an almost quantitative agreement. For these three observables, the biggest discrepancy is observed at low temperatures for alkanes with a chain longer than 10 carbon atoms, with the formation of smectic layers and a solid phase. The computed viscosity is in good agreement with the experimental data for butane and hexane. Significant positive deviations are observed for the other alkanes, with the largest error being three orders of magnitude for hexadecane. The observed discrepancy increases with the chain length, suggesting that the source of error is an extensive quantity. A first possible source of error could be that the bond potential between beads is biased towards the formation of the

extended configuration, therefore driving the formation of ordered phases due to reduced conformational disorder.

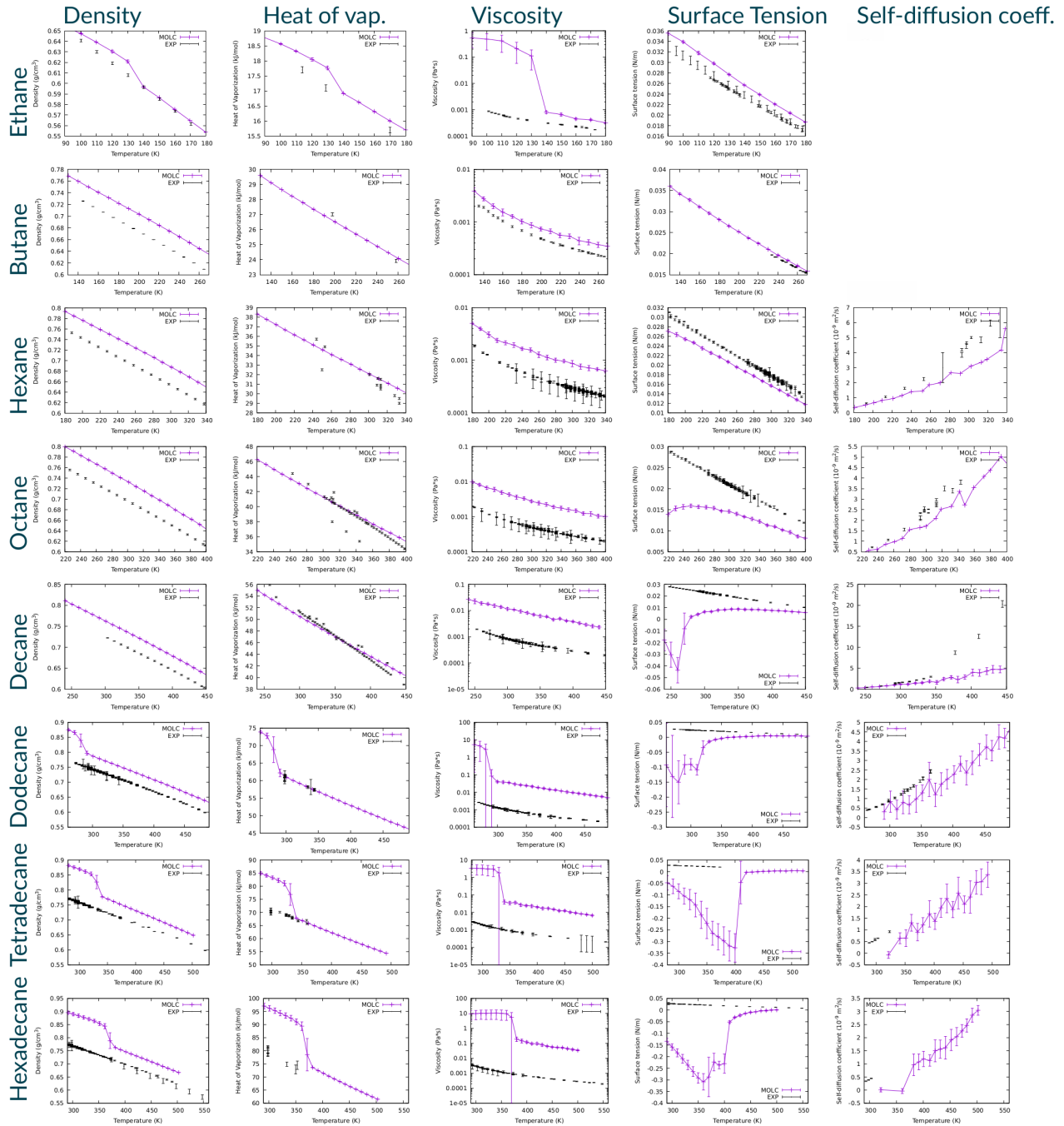


Figure 4: Physical observables computed for linear alkanes: density (g/cm^3), enthalpy of vaporisation (kJ/mol), viscosity ($\text{Pa}\cdot\text{s}$), surface tension (N/m), and self-diffusion coefficient (m^2/s).

This effect could be seen as the opposite of lowering the melting point in unsaturated lipids. Another source of error is that the CG model has an intrinsically larger moment of inertia due to spreading the mass on finite-size particles instead of being concentrated on point particles. For instance, the diagonal inertia tensor of the ethane molecule is (6.5, 25.4, 25.4) $\text{amu}\cdot\text{\AA}^2$, while that of the corresponding CG bead is (45.1, 62.3, 62.3) $\text{amu}\cdot\text{\AA}^2$.

The surface tension agrees with the experimental data for chain lengths up to 8 carbon atoms. For heavier alkanes, increasingly larger deviations are observed at low temperatures due to the formation of an ordered interface with the vacuum, as shown in Fig. 5. The observed error could be due to the finite-size size of the sample, which is amplified for longer alkanes. Since 1000 molecules were used in all simulations, a comparatively larger fraction of molecules are segregated in the interface for samples of the heavier alkanes compared to the bulk of the free-standing film. In addition, the considerations on the CG model done for explaining the discrepancy observed on the computed viscosity apply here too. All the considerations made to explain the observed discrepancies can be systematically tested with new simulations. This work is outside the scope of this project but is reported here for the sake of discussion.

The effect of the bond potential can be explored by replacing an alkyl bead with one describing a cis-alkene, which introduces a kink along the chain and should prevent the tails from packing closely together. The expected outcome is to increase the conformational disorder and to make the model less likely to form a solid. The beads' size and partial overlap along the chain can introduce spurious 1-3 interactions, which may provide an extra energy stabilisation for the extended chain configuration compared to a more disordered structure. This effect could be easily investigated by setting 1-3 and 1-4 weighting coefficients to zero. The impact of the larger inertia tensor for the CG model could be investigated by simply scaling down the masses to minimise the difference with the atomistic model.

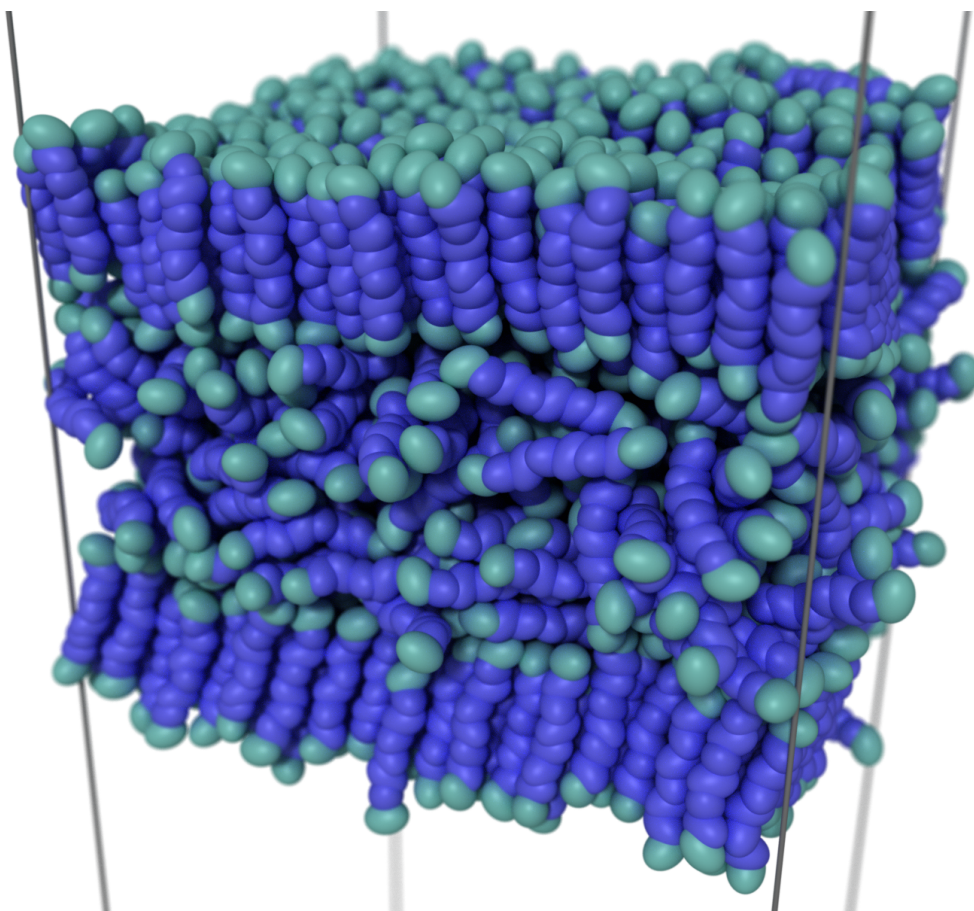


Figure 5: A free-standing film of 1000 hexadecane molecules at 360 K, showing the formation of an orderer interface with the vacuum.

In addition to the described work on linear alkanes, part of the computing time was used to parametrise the poly lactic-co-glycolic acid (PLGA) polymer to test the `compute intra` command for both the atomic and coarse-grained models of PLGA. An initial guess of the CG parameters was produced using acetic acid and isobutyric acid as “effective” monomers since they correspond to the repeating units along the PLGA chain. With this initial set, a CG sample of PLGA was produced and back-mapped to the atomic structure to produce reference data in the form of mass density and intermolecular energy. The GB parameters of the CG model were then refined using these reference values, yielding good agreement in terms of both structural (density) and mechanical (elastic moduli) properties.

Computational efficiency

The present work has been carried out using a new pair style that replaces the previous implementation of the MOLC model, based on overlaying a standard Gay-Berne potential with a custom Coulombic pair style where the charges are defined in the frame of reference of the parent ellipsoid¹. The new pair potential was implemented by Stephen Farr in the LAMMPS development branch, version 2 June 2022. The new source code is:

```
pair_molc_cut.cpp/.h
pair_molc_long.cpp/.h
pppm_molc.cpp/h
```

The usage for the truncated style is:

```
# gamma epsilon mu lj_cutoff coul_cutoff
pair_style molc/cut 1 1 -3 12.0 12.0

# Pair coefficients.
# eps0, sigma0, eps_x, eps_y, eps_z, nq (xi, yi, zi, qi)_nq times
pair_coeff 1 1 0.102 3.188 1 1 1 3 &
0.000000 0.000000 0.000000 -0.830 &
-0.756950 0.585882 0.000000 0.415 &
0.756950 0.585882 0.000000 0.415
```

The usage for the style with a long-range solver is:

```
# gamma epsilon mu lj_cutoff coul_cutoff
pair_style molc/long 1 1 -3 12.0 12.0

# no longer need to additionally specify the charges here
kspace_style ppm/molc 1e-4

# Pair coefficients.
# eps0, sigma0, eps_x, eps_y, eps_z, nq (xi, yi, zi, qi)_nq times
pair_coeff 1 1 0.102 3.188 1 1 1 3 &
0.000000 0.000000 0.000000 -0.830 &
-0.756950 0.585882 0.000000 0.415 &
0.756950 0.585882 0.000000 0.415
```

The new MOLC pair style was verified by comparing the computed energy and forces of reference systems with those of the former implementation. Using a single pair style instead of overlaying more pair potentials should improve the computational performance since a single neighbour list is used. The new MOLC pair style is faster than the old model, with the most significant gains being for the pair molc/cut style, while the computational cost of the pair molc/long style is dominated by the K-space term. The tables below show the performance difference for the new MOLC pair style for different core counts and test systems.

TIP4P/2005, pair style: molc/cut.

Nodes	Old steps/s	New steps/s	Percentage change
1	23.102	26.391	14.2
2	46.417	52.19	12.4
4	87.229	97.995	12.3
8	166.891	189.897	13.8
16	277.968	299.356	7.7
32	439.11	477.853	8.8

α -NPD, pair style: molc/cut.

Nodes	Old steps/s	New steps/s	Percentage change
1	3.458	3.626	4.9
2	6.802	7.153	5.2
4	13.312	14.132	6.2
8	25.471	27.138	6.5
16	50.28	53.084	5.6
32	96.754	103.308	6.8
64	164.995	172.386	4.5
128	311.371	320.021	2.8

TIP4P/2005, pair style: molc/long.

Nodes	Old steps/s	New steps/s	Percentage change
1	18.547	19.479	5.0
2	36.276	37.994	4.7

4	67.069	70.337	4.9
8	123.092	123.092	3.7

Conclusions

In this project, the pair style for the MOLC coarse-grained model was ported from an old version of LAMMPS to the latest development version. The code was optimised for using a single list of neighbour molecules, leading to a performance improvement between 3 and 14%. In addition, a new command for computing the inter-molecular energy was developed and tested on different systems. These improvements resulted in a more robust parametrisation strategy, which was validated for linear alkanes. The results showed almost quantitative agreement between the coarse-grained simulations and reference experimental data for several physical observables. The outcomes of this project have greatly advanced the accuracy and efficiency of the modelling framework developed by MaterialX. The work presented here will be published in a peer-reviewed paper and used to model lipid and surfactant systems in subsequent work.

Acknowledgements

This work was financially supported by the PRACE project. The authors acknowledge the SHAPE (SME HPC Adoption Programme in Europe) 14th programme for granting access to the Archer2 supercomputing facilities.

References

1. M. Ricci, O. M. Roscioni, L. Querciagrossa, C. Zannoni, *Phys. Chem. Chem. Phys.*, **21**, 26195 (2019).
2. M. Allen and D. Tildesley, *Computer simulation of liquids* (Oxford Science, 1989).
3. N. Karasawa and W. A. Goddard, *J. Chem. Phys.*, **93**, 73207327 (1989).
4. A. Y. Toukmaji and J. A. Board. *Comput. Phys. Commun.* **95**, 73–92 (1996). DOI: 10.1016/0010-4655(96)00016-1
5. J. Wang and T. Hou, *J. Chem. Theory Comput.* **7**, 2151–2165 (2011).
6. Diky V.; Chirico, R.D.; Lemmon, E.W.; Muzny, C.D.; Kazakov, A.F.; Kroenlein, K.; Magee, J.W.; Abdulagatov, I.; Kang, J.W.; Frenkel, M. *J. Chem. Inf. Model.* **53**, 3418–3430 (2013).
7. S. J. Marrink, A. H. de Vries, A. E. Mark, *J. Phys. Chem. B*, **108**, 750–760 (2004).
8. C. Avendaño, T. Lafitte, C. S. Adjiman, A. Galindo, E. A. Müller, G. Jackson, *J. Phys. Chem. B*, **117**, 2717–2733 (2013).
9. Y. An, K. K. Bejagam, S. A. Deshmukh, *J. Phys. Chem. B*, **122**, 7143–7153 (2018).
10. J. J. P. Stewart, *J. Comput.-Aided Mol. Des.*, **4**, 1–103 (1990).
11. A. K. Malde, L. Zuo, M. Breeze, M. Stroet, D. Poger, P. C. Nair, C. Oostenbrink, A. E. Mark, *J. Chem. Theory Comput.*, **7**, 4026–4037 (2011).