Development of Machine Learned Potentials for Coarse Grained Simulations of Polymers

Eleonora Ricci
National Centre for Scientific Research "Demokritos"
Athens, Greece
e.ricci@inn.demokritos.gr

George Giannakopoulos
National Centre for Scientific Research "Demokritos"
Athens, Greece
ggianna@iit.demokritos.gr

Vangelis Karkaletsis
National Centre for Scientific Research "Demokritos"
Athens, Greece
vangelis@iit.demokritos.gr

Doros N. Theodorou
National Technical University of Athens
Athens, Greece
doros@chemeng.ntua.gr

Niki Vergadou
National Centre for scientific Research "Demokritos"
Athens, Greece
n.vergadou@inn.demokritos.gr

ABSTRACT

In this work, we adopted Graph Convolutional Neural Networks (GCNN) to develop Coarse Grained (CG) Machine Learned potentials for bulk polymer systems, implementing a scheme that includes a force-matching procedure. The GCNN models were used to perform CG Molecular Dynamics (MD) simulations of polyethylene and of a polymer of intrinsic microporosity (PIM-1). The structural and thermodynamic properties of the CG systems were compared with the underlying atomistic reference, examining the effect of the CGNN model size and hyperparameters on the simulation results. The models obtained showed transferability to longer chain lengths than the ones use for training. The open-source python package SchNetPack was extended to support the study of macromolecular systems with connectivity information and inter- and intra-molecular neighbours distinctions. The extended version is are available on GitHub at the following link: https://github.com/ml-multimem/schnetpack-for-bulk-systems-for-bult-systems. Moreover, the GCNN models were interfaced with the MD engine LAMMPS. This methodology has the potential to streamline the generation of CG force fields, enabling systematic multiscale studies of complex macromolecular systems.

1. Introduction

Polymers play a fundamental role in a variety of processes and novel technologies, such as manufacturing, biomedical, energy, and environmental applications, owing to their versatile properties, originating from their diversity in chemical structure and macromolecular architecture. However, the design of task-specific materials with controlled properties is a lengthy and resource-intensive trial-and-error process. Rational and targeted materials design necessitates a fundamental understanding of the structure-properties link, and an unravelling of the microscopic mechanisms that control end-use performance at the relevant operating conditions. Simulation and modelling studies play a key role in filling this gap, advancing fundamental understanding and accelerating technology transfer [1]. However, several unsolved challenges limit the application of molecular simulation methods to macromolecules of complex chemical constitution, such as high molecular weight (M_w) polymers, especially at bulk conditions. A comprehensive study of these systems necessitates the development of innovative multi-scale simulation techniques [2,3].

Coarse-Graining (CG) [4] is, in many cases, at the core of multiscale methods, especially for the case of polymeric systems. Coarse Graining consists of averaging out a certain number of high-resolution degrees of freedoms, namely atoms, into a single interaction site, or coarse-grained bead (*i.e.* choosing a CG mapping), and represent the interactions between the CG beads consistently with the underlying atomistic system, namely developing a CG force field, or CG potential. Adopting a CG molecular description enables access to larger lengths and longer timescales, thereby expanding the reach of systems and phenomena that can be investigated with molecular modelling.

Traditional hierarchical simulation methods and models [5] typically use predefined pairwise functionals to describe particle interactions, which inherently limits their ability to describe many-body interactions. In order to overcome the limitations of traditional methods, in this work we investigated the development of an ML-based hierarchical simulation strategy for bulk macromolecular systems, bridging atomistic and coarse-grained scales of description. ML-based strategies have been applied to bridge the quantum-mechanical and atomistic levels of descriptions, mostly for the case of inorganic systems or isolated small organic molecules [6]. However, their application to complex macromolecular systems at bulk conditions has been scarcely explored so far [7]. In particular, in this work we explored the possibility to train ML models, such as neural networks, to replace traditional force fields based on predefined functionals.

We considered a Graph Convolutional Neural Network (GCNN) architecture named SchNet [8], which had previously shown promise in the development of atomistic force fields, trained on quantum mechanical calculations. We tested the approach on a benchmark polymer, i.e. polyethylene, considering 2 different CG mappings, as shown in Figure 1, as well as on a polymer of intrinsic microporosity (PIM-1). PIM-1 is a material of great interest for environmental applications such as gas separation with membranes. It represented a significant breakthrough in terms of separation performance and its complex chemical structure makes it a challenging system to simulate with conventional methods [9].

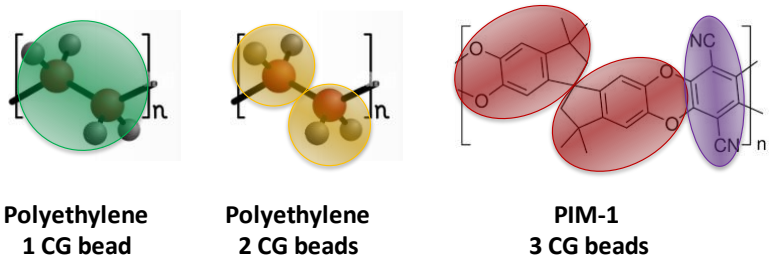

Figure 1: Systems studied for the development of GCNN force fields, with their respective CG mappings.

To train ML potentials and perform CG simulations, we extended the open-source python package SchNetPack [10,11] to enable the support for systems including connectivity information and inter- and intra-molecular neighbours distinctions, which are key aspects for the description of macromolecular systems at bulk conditions. Moreover, the approach is integrated within the Molecular Dynamics engine LAMMPS [12], which is one of the most widespread codes in use in the molecular simulation community.

2. Methods

2.1 Force Matching

A force matching scheme [13] was adopted for the determination of the CG potential. This method involves the minimization of the difference between the atomistic forces, $\mathbf{F}^{\mathrm{A}} \in \mathbb{R}^{3n}$ (where n is the number of atoms), projected on the CG sites, and the forces predicted by the CG model, $\mathbf{F}^{\mathrm{CG}} \in \mathbb{R}^{3N}$, for the N number of CG sites having coordinates $\mathbf{x}_i$:

$$\chi^2 = \langle \left[\mathbf{F}^{\mathrm{CG}}(\mathbf{x}_i) - \mathcal{M}\left(\mathbf{F}^{\mathrm{A}}(\mathbf{r}_j)\right) \right]^2 \rangle \qquad (1)$$

$\mathcal{M}$ is the mapping operator between the atomistic and the CG space, $\mathbf{r}_j$ are the atomistic coordinates and the brackets $\langle \ldots \rangle$ denote average over CG degrees of freedom and the number of sampled atomistic configurations. The force matching approach does not include structural information, such as radial distribution functions, during the fitting, therefore, the ability of a CG model to represent the structure is a true test of quality and an important characteristic that can be used for the comparison among different models.

During the training, the mean squared error between the predicted forces and the ground truth values of the forces, associated with the molecular dynamics (MD) trajectory considered as input, is minimized, using the following loss function L:

$$L = \frac{1}{3\mathrm{N}} \sum_i^{\mathrm{N}} \left\{ \left[-\nabla_{\mathbf{x}_i} (\Sigma_i \widehat{U}_i + U_{ex} + U_{bond}) \right] - \mathcal{M}\left(\mathbf{F}^{\mathrm{A}}(\mathbf{r}_j)\right) \right\}^2 \qquad (2)$$

where N is the total number of CG particles, $\widehat{U}_i$ are the per-particle energy values predicted by the CGNN, $\mathcal{M}\left(\mathbf{F}^{\mathrm{A}}(\mathbf{r}_j)\right)$ is the force acting on the CG particle at position $\mathbf{x}_i$, obtained from the atomistic MD simulation mapped to the CG space, while the term in square brackets is the force acting on the CG particle at position $\mathbf{x}_i$ predicted by the GCNN. The term U_{ex} is an excluded volume prior energy term based on the pairwise distances between the moieties, included to ensure that the energy will become steeply very high when the configuration assumed by the system is departing from physical states, for example when atoms are moving too close to each other [14]:

$$U_{ex} = \sum_{i=1}^{N-1} \sum_{j=i}^{N} \left(\frac{\sigma}{\|x_i - x_j\|} \right)^{n^{ex}} \qquad (3)$$

σ and n^{ex} are hyperparameters, namely the excluded volume diameter and the excluded volume exponent, for which suitable values must be identified. Connectivity information was included into the model in the form of a harmonic bond prior energy term:

$$U_{bond} = k_b (l - l_0)^2 \qquad (4)$$

Where l_0 is the equilibrium CG bond length, l is the actual bond length, and k_b is the bond stiffness constant, which in the present case was considered as a model hyperparameter. Analogously to the case

of the excluded volume term, the addition of this prior term ensures that the energy will become steeply very high when unrealistically stretched or compressed bonds occur in the system.

Previous work [15,16] suggested that it can be beneficial to modify the loss function including also the squared deviation between the energy value predicted by the model and an energy component from the atomistic simulation (U_A), appropriately weighted (λ) with respect to the force component. Thus, the loss function definition becomes:

$$L = \frac{1}{3N}\sum_i^N \left(-\nabla_{x_i}(\Sigma_i \widehat{U}_i + U_{ex} + U_{bond}) - \mathcal{M}\left(\boldsymbol{F}^A(\boldsymbol{r_j})\right)\right)^2 + \lambda\left(\left(\Sigma_i \widehat{U}_i + U_{ex} + U_{bond}\right) - U_A\right)^2 \quad (5)$$

2.2 SchNet Graph Convolutional Neural Network Architecture

The main features of the ML architecture utilized are summarized hereafter. For a more detailed description, the original works can be consulted [8,17]. The GCNN architecture utilized is named SchNet and it adopts a graph representation of molecules, with nodes corresponding to particles and edges to bonds or interatomic distances. Each particle is represented through a feature vector, which is initialized to distinguish between the particle chemical identities (embedding layer). The feature is then updated to encode information on the surrounding environment, by performing continuous convolutions across the particle neighbourhood, optimizing the convolutional filter weights during the training (convolutional layers). By performing multiple convolution operations, each node can influence other increasingly distant nodes, thereby potentially encoding long-range information into the particle descriptor. Afterwards, a fully connected section is included (readout layers). A sequence of continuous convolution layers and readout layers constitutes a SchNet "block". Multiple blocks can be utilized in series to define the full network architecture. The output found at the end of the last block can be interpreted as a learned feature representation, which encodes the information from the particle neighbourhood required to predict the target property. Finally, given the learned feature representation as input, a fully connected (dense) section is tasked with predicting the final scalar output, which is interpreted as a per-particle energy contribution, $\widehat{U}_i$. This local decomposition ensures invariance to the total number of particles. All energy contributions are summed to obtain the total energy ($\Sigma_i \widehat{U}_i$), which is then differentiated with respect to the positions of the particles to predict the force acting on each particle. In this way, a conservative force field is obtained by design.

Overall, there are several hyperparameters to optimize in the SchNet architecture, and this proved to be the main hurdle encountered in the application of this method. Hyperparameters are user-defined settings of ML models that control various aspects of the model architecture and behaviour, such as (in this case):

(a) the model size & architecture (number of convolutional filters, filters size, number of SchNet blocks, activation function)

(b) the particle representation (feature vector size, cutoff radius of the local neighbourhood)

(c) the training process (learning rate, decay ratio, batch size, number of samples, number of epochs)

(d) the embedded prior terms (excluded volume diameter, excluded volume exponent)

2.3 Extensions to the SchNetPack python package

We extended the open-source python package SchNetPack [10,11] to enable ML CG force field development for multi-molecule and macromolecular systems. Originally, SchNetPack had been created to develop atomistic ML force fields from quantomechanical simulations. A new "*Particle*" class was defined, which contains additional properties compared to the primitive "Atom" one: *particle type*, *molecule membership*, *connectivity information*. Various neighbours' list construction criteria were implemented, to enable to application of selected features of the model architecture only to a certain subset of neighbours, such as the bonded/nonbonded or inter-/intramolecular ones. Several other classes were extended to support these new functionalities for training and simulation with the ASE python package [18]. The possibility to include prior energy terms to the output modules was added, and excluded volume and harmonic bond prior classes were implemented. We extended also the corresponding SchNetPack-LAMMPS interface, for the utilization of the trained GCNN force fields within this widespread and highly optimized MD engine. The code developed in this work is available on GitHub at the following link: https://github.com/ml-multimem/schnetpack-for-bulk-systems.

2.4 Dataset generation, training and simulation details

The training data consisted of atomistic MD trajectory frames. Each system contained 10 polymer chains, of 50 monomers in the case of polyethylene, of 10 monomers in the case if PIM-1. They were simulated at 300 K and atmospheric pressure, under periodic boundary conditions. The polyethylene

system was initially equilibrated through a 1 ns NVT MD simulation, then 20 ns NPT, and then 20 ns NVT simulation at the average equilibrium density, 0.843 g/cm^3. The PIM-1 system was initially equilibrated through a 1 ns NVT MD simulation, then simulated using the 21 step equilibration cycle described by Colina and coworkers [19], increasing the duration of each step by a factor 10, following with a 20 ns NPT simulation at atmospheric pressure, and then a 20 ns NVT run at the average equilibrium density of 1.069 g/cm^3. For both materials, the first 10 ns of the last NVT run were discarded, and, after that, 10000 configurations were retained, using 9000 for the training set and 1000 for the test set. The molecular configurations from the MD trajectory were randomly assigned to either set, and shuffled during the training. Simulations of the atomistic system were conducted with the pcff force field [20], using LAMMPS [12]. LAMMPS input files to perform the simulations that were used as training sets can be downloaded at the following link: https://zenodo.org/records/10352368.

After the model training, subsequent NVT CG simulations with the ML CG potentials were conducted using the Atomic Simulation Environment (ASE) python package [18]. The loss function definition reported in Eq. 5 was used for the training, either considering both energy and forces or only forces (by setting $\lambda = 0$). In the first case, different values for the target energy value U_A were tested (listed in Table 1) and the best one was found to be the sum of the intermolecular and bond energy contributions from the atomistic simulation. Table 1 summarizes the ranges investigated for each hyperparameter for both polymers, for a total of approximately 50 combinations for each material investigated.

For polyethylene, two different mappings were considered: 2 CG beads per monomer, in which each bead represents one carbon atom and its bonded hydrogens (often referred to as “united atom” mapping), and 1 CG bead per monomer, in which 2 carbon atoms and their bonded hydrogens make up one bead. These systems contain 1000 and 500 interaction sites, respectively. The chain ends were distinguished from the backbone beads when learning the feature vector, by assigning them a different embedding in the SchNet embedding layer. PIM-1 is a copolymer obtained from 2 precursors: 3,3,3′,3′-tetramethyl-1,1′-spirobiindane-5,5′,6,6′-tetraol, and 2,3,5,6-tetrafluorophthalonitrile. 2 CG beads were used to represent the section of the monomer resulting for the first precursor (shown in red in Figure 1), and 1 CG bead was used for the section of the monomer resulting from the second precursor (shown in violet in Figure 1). So overall, 3 CG beads per monomer were used, resulting in 300 CG interaction sites. The two “red” beads are almost identical: one contains the carbon atom situated in the spirobiindane contortion site, and therefore has a higher mass than the other. A different type embedding was therefore

used for each of the 3 beads. Moreover, the chain ends were distinguished from the backbone beads, leading to 5 different bead types in the embedding layer of the network. For both polymers, in the prior energy terms the same values for the hyperparameters σ, n^{ex}, and k_b were used for all bead types, to simplify the hyperparameter optimization process.

Table 1. SchNet hyperparameter ranges investigated.

Polyethylene			
Hyperparameter	**Values**	**Hyperparameter**	**Values**
Excluded volume radius (σ) [Å]	1.5, 2, 2.5, 3, 4	Number of SchNet blocks	2, 3, 4
Excluded volume exponent (n^{ex})	5	Target energy	Intermolecular, intermolecular + bonded
Local neighbourhood cutoff radius [Å]	6, 7, 8	Bond stiffness k_b [$\frac{kcal}{mol\ Å^2}$]	20, 100, 200, 1000
Energy scaling in the loss function (λ)	0, 0.01, 0.1, 0.5, 1	Batch size	20, 50
PIM-1			
Hyperparameter	**Values**	**Hyperparameter**	**Values**
Excluded volume radius (σ) [Å]	4, 5, 6, 7, 8, 9	Number of SchNet blocks	2, 3
Excluded volume exponent (n^{ex})	5, 6	Target energy	Total potential, intermolecular, intermolecular + bonded
Local neighbourhood cutoff radius [Å]	8, 9, 10, 15	Bond stiffness k_b [$\frac{kcal}{mol\ Å^2}$]	20, 50, 100, 200
Energy scaling in the loss function (λ)	0, 0.01, 0.05, 0.1, 0.5, 1	Batch size	10, 50, 100

3. Results

The training of a successful model proved to be a computationally intensive task, primarily due to the extensive testing required to identify a hyperparameter combination that would yield satisfactory simulations in terms of stability and structural and dynamic representation. In the majority of cases, very similar final values of the loss function were reached when using different hyperparameter

combinations. However, when the models were used for simulations, very different and at times unphysical behaviours were observed. Similar challenges were described also for other systems [15,16,21,22]. The discrepancy between the metric employed during training (mean square error on the prediction of the forces) and the criteria used to assess a model's ability to produce meaningful simulation results (mainly related to thermodynamic stability and accurate reproduction of structural characteristics) posed a significant challenge. It was unclear how one could leverage simulation results to inform the training process and guide hyperparameter search. As a consequence, the exploration of the hyperparameters space became a time and resource-intensive trial-and-error process. In the polymer systems investigated, simulation stability issues tended to manifest less frequently compared to other cases, such as bulk liquid systems [15], therefore the hyperparameter search was mostly focused on the structural representation, evaluated by computing the radial distribution function, or g(r), between the centres of mass of the CG beads, and its intra- and intermolecular components.

The best model obtained for each material, as well as full specifications of their hyperparameters, can be found at the following link: https://zenodo.org/records/10343599.

3.1 Results obtained for polyethylene

The best results obtained for the case of polyethylene mapped using 2 CG beads are shown in Figure 2. It was observed that a very good structural representation could be achieved when training the model on the force differences alone, i.e. setting the scaling factor $\lambda = 0$. Irrespective of the scaling factor chosen, when the energy component was considered in the loss function, the intermolecular component of the radial distribution function would always display too weak structural features compared to the atomistic reference. The excluded volume radius did not influence this behaviour within the range investigated. The bond stiffness constant did not affect the structure appreciably as well, but we noted that very high values tended to favour instability. The best result was obtained considering σ = 1.5 Å, $\lambda = 0$, cutoff = 7 Å, $k_b = 200 \frac{kcal}{mol\ Å^2}$.

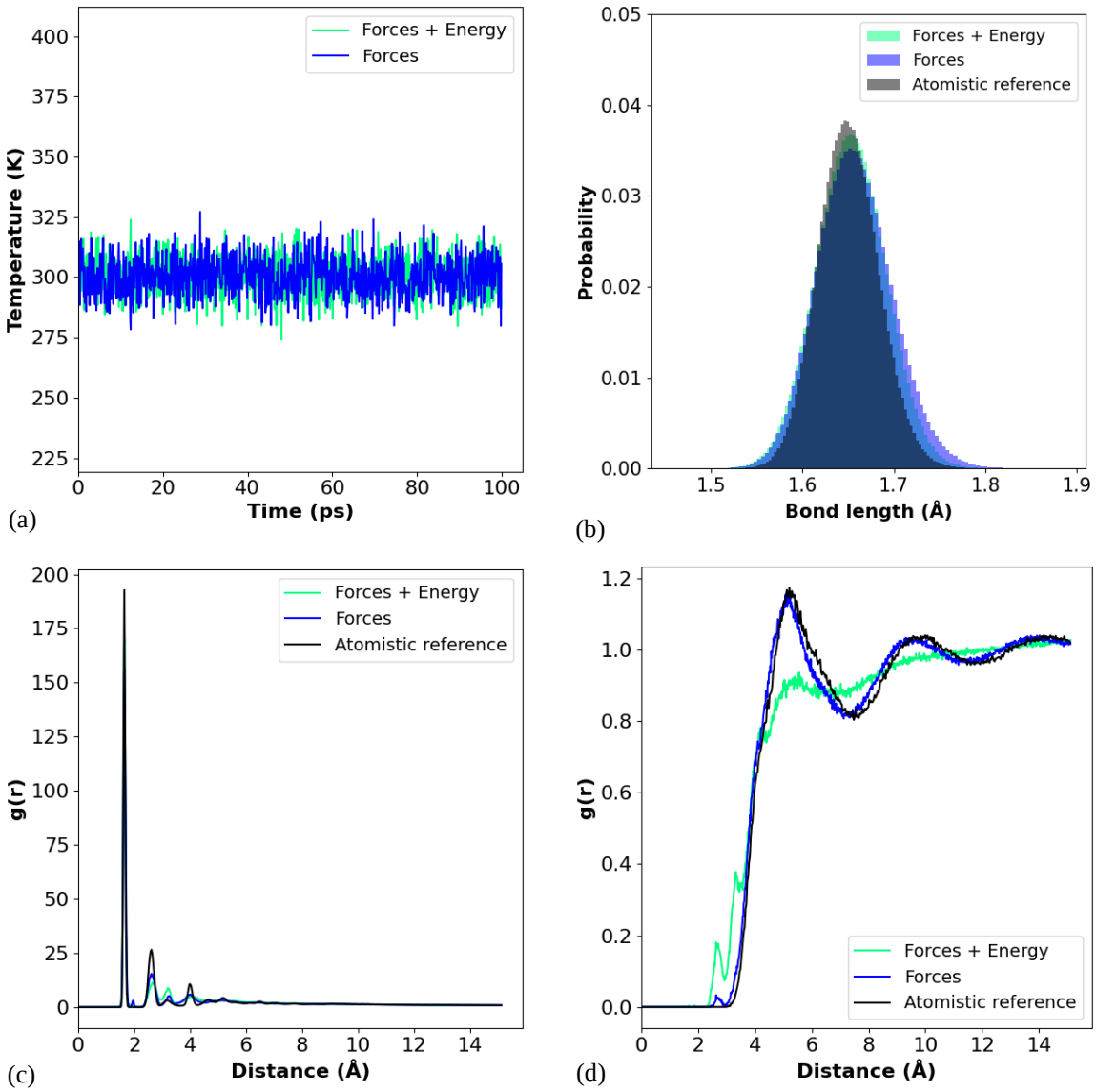

Figure 2: Comparison of the (a) temperature trend, (b) backbone bonds length probability distribution, (c) intramolecular radial distribution function, and (d) intramolecular distribution function during CG NVT MD simulations of polyethylene mapped using 2 CG beads per molecule, performed at 300 K and atmospheric pressure with GCNN potentials trained using Eq. 5 as the loss function, with different values of the energy scaling factor λ (blue: λ=0, green: λ=0.1, black atomistic reference).

In the case of polyethylene mapped using 1 CG bead per monomer, the match between the reference intermolecular g(r) and the one obtained from the CG MD simulations is not as close as in the case of the united atom system, as shown in Figure 3. Nonetheless, also in this case, a better result was obtained considering only force differences in the loss function. Several attempts were made to find a

combination of hyperparameters that would yield an improved structural representation. The best result, shown in Figure 3, was obtained considering σ = 4 Å, λ = 0, cutoff = 7 Å, k_b = 20 $\frac{kcal}{mol\ Å^2}$.

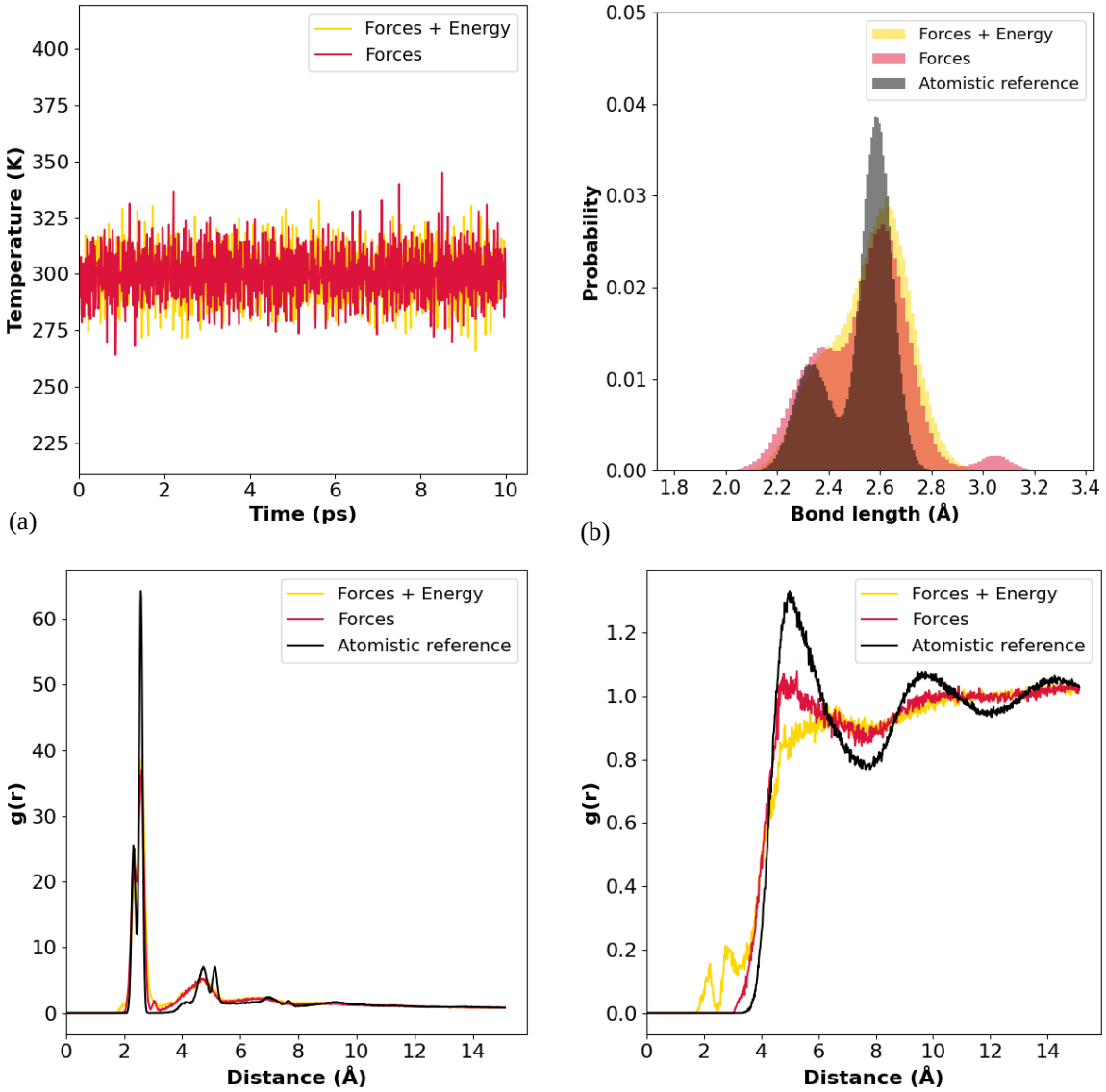

Figure 3: Comparison of the (a) temperature trend, (b) backbone bonds length probability distribution, (c) intramolecular radial distribution function, and (d) intramolecular distribution function during CG NVT MD simulations of polyethylene mapped using 1 CG bead per molecule, performed at 300 K and atmospheric pressure with GCNN potentials trained using Eq. 5 as the loss function, with different values of the energy scaling factor λ (red: λ=0, yellow: λ=0.5, black atomistic reference).

CG MD simulations of longer chain systems were performed in both CG representations, utilizing the best models obtained. Double and quadruple M_w were considered, simulating two systems containing 5 chains of 100 monomers and 3 chains of 200 monomers respectively. As can be seen in Figure 4, the systems exhibited consistent thermodynamic, structural, and dynamics characteristics, thus giving positive indication about the size transferability of the GCNN force fields to higher M_w polymer chains.

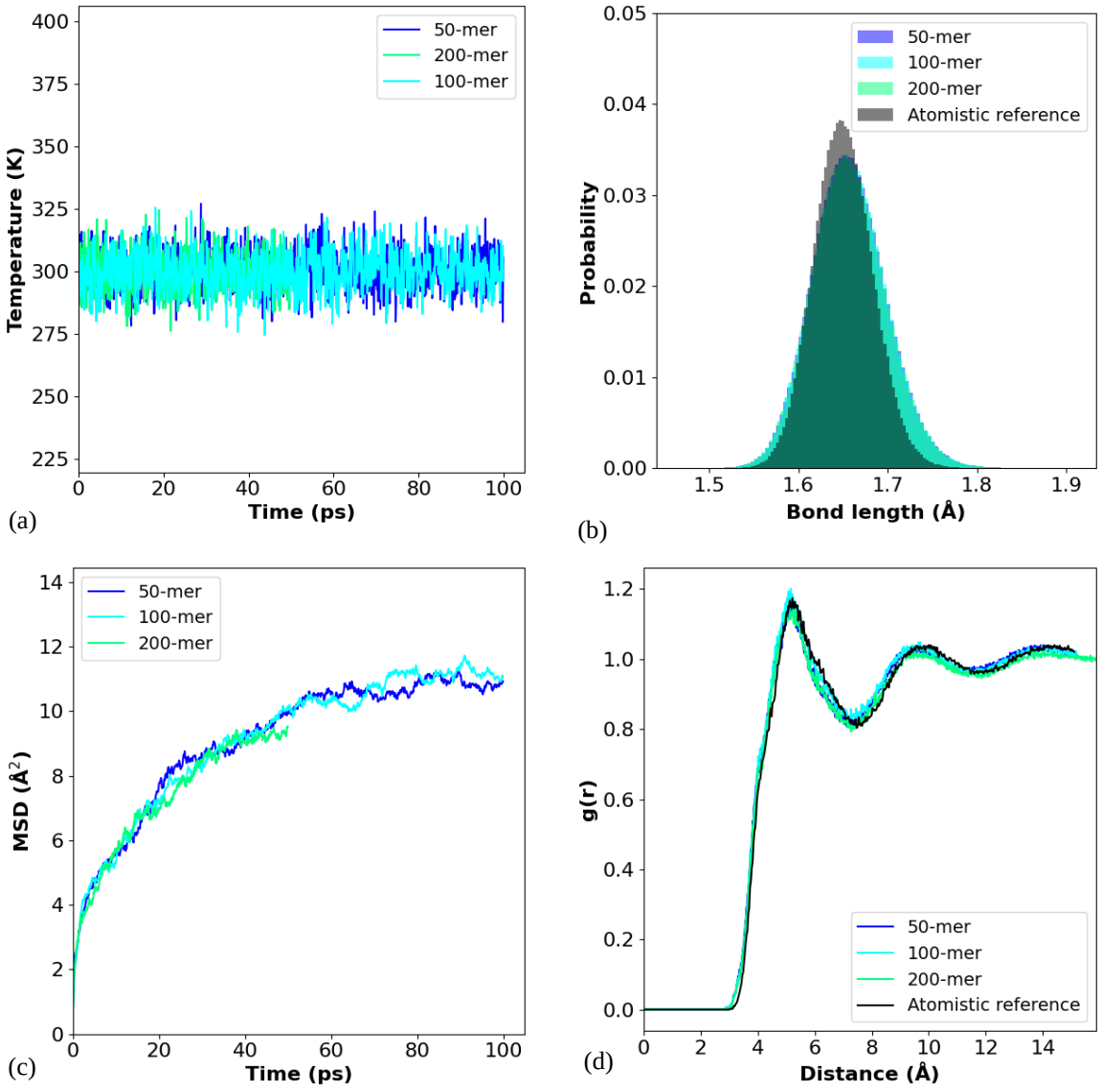

Figure 4: Comparison of the (a) temperature trend, (b) bond length probability distribution for one of the backbone bond types (bond between a "red" and a "violet" bead – see Figure 1), (c) Mean Square Displacement of the centre of mass of the CG beads, and (d) intramolecular distribution function during CG NVT MD simulations of PE of different molecular weight mapped using 2 CG beads per molecule, performed at 300 K and atmospheric pressure with a GCNN potential trained on 50-mers.

3.2 PIM-1

Figure 5 shows the best results obtained for PIM-1 considering only force differences in the loss function, or including also the energy contribution, appropriately weighted. For this polymer, a better structural representation was observed when the energy contribution was considered, both concerning the inter- and intramolecular parts of the g(r). The bond length probability distributions were more spread than the atomistic reference one, and it was observed that increasing the value of k_b had a slightly positive effect, but if too high values were used, the system would freeze.

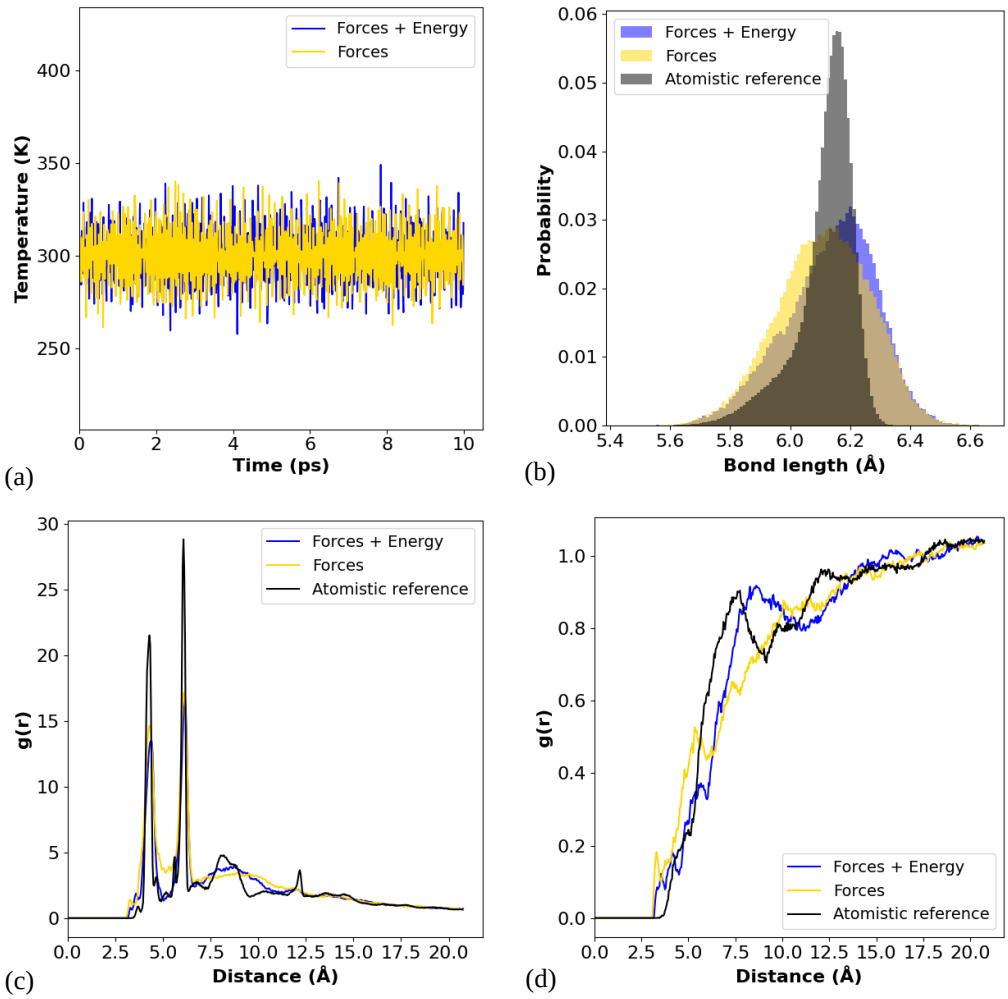

Figure 5: Comparison of the (a) temperature trend, (b) bond length probability distribution for one of the backbone bond types (bond between a "red" and a "violet" bead – see Figure 1), (c) intramolecular radial distribution function, and (d) intramolecular distribution function during CG NVT MD simulations of PIM-1 mapped using 3 CG beads per molecule, performed at 300 K and atmospheric pressure with GCNN potentials trained using Eq. 5 as the loss function, with different values of the energy scaling factor λ (blue: λ=0.05, yellow: λ=0, black atomistic reference).

The effect of σ was also investigated and can be seen in Figure 6. It did not affect the stability of the simulations, but high values seem to result in slower dynamics of the system and are also associated to the appearance of spurious peaks in the g(r), both inter- and intramolecular. It is worth noting that, in this system, the backbone is made of beads of different mass and representing different underlying chemical identities, contrary to the case of polyethylene, where all backbone beads are identical among themselves and minimally different from chain-end beads. For this reason, considering the same values for the energy prior hyperparameters could introduce a greater level of approximation in this system. This was not investigated at this stage, but it is noted for future work.

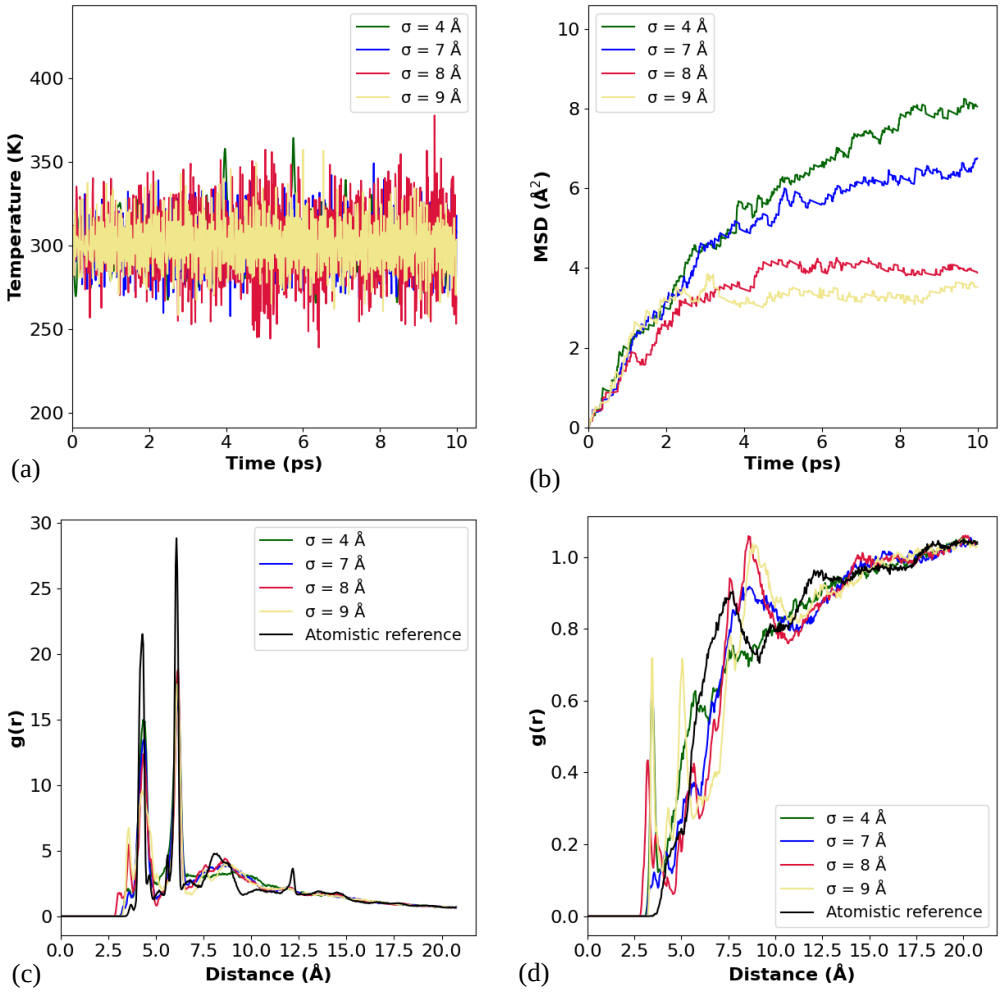

Figure 6: Comparison of the (a) temperature trend, (b) MSD of the centre of mass of the CG beads, (c) intramolecular radial distribution function, and (d) intramolecular distribution function during CG NVT MD simulations of PIM-1 mapped using 3 CG beads per molecule, performed at 300 K and atmospheric pressure with GCNN potentials trained using Eq. 5 as the loss function, with energy scaling factor λ=0.05, and various values of the excluded volume diameter σ (green: σ = 4 Å, blue: σ = 7 Å, red: σ = 8 Å, yellow: σ = 9 Å, black atomistic reference).

An interesting effect associated to the value of the local neighbourhood cutoff radius was observed. Increasing the cutoff lead to a more pronounced overfitting behaviour, as shown in Figure 7: the energy component of the loss would decrease more markedly, and, concurrently, the force component would experience an increase. The model with the overall lowest value of the loss function over the validation set was used for the subsequent simulation, which for the case of Figure 7a was reached at epoch 56. During the simulations, the three models differed mostly in the intermolecular structural representation, and, to a lower extent, also in the intramolecular one, as shown in Figure 8.

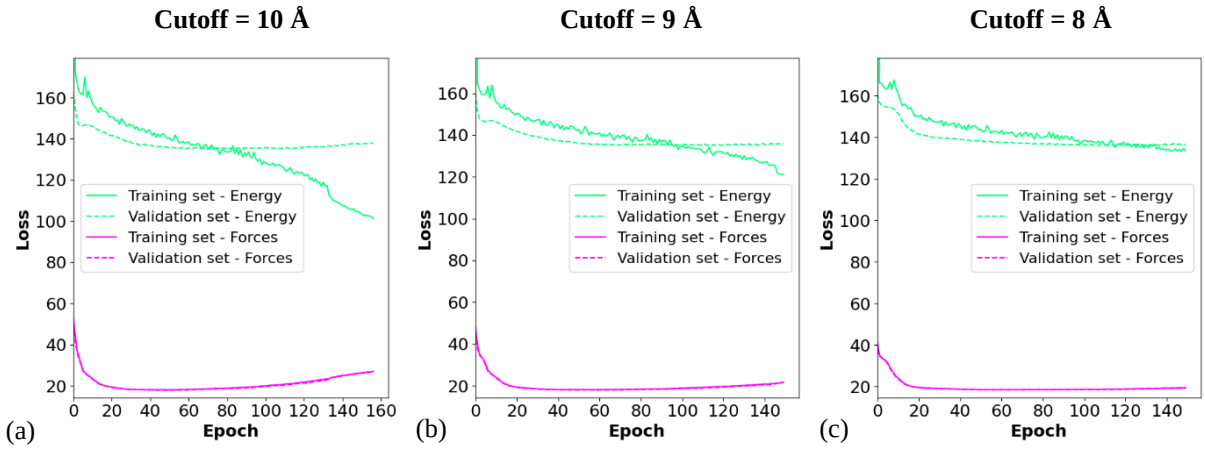

Figure 7: Comparison of the trends of the loss components calculated over the training and validation sets for models trained with different values of the local neighbourhood cutoff radius.

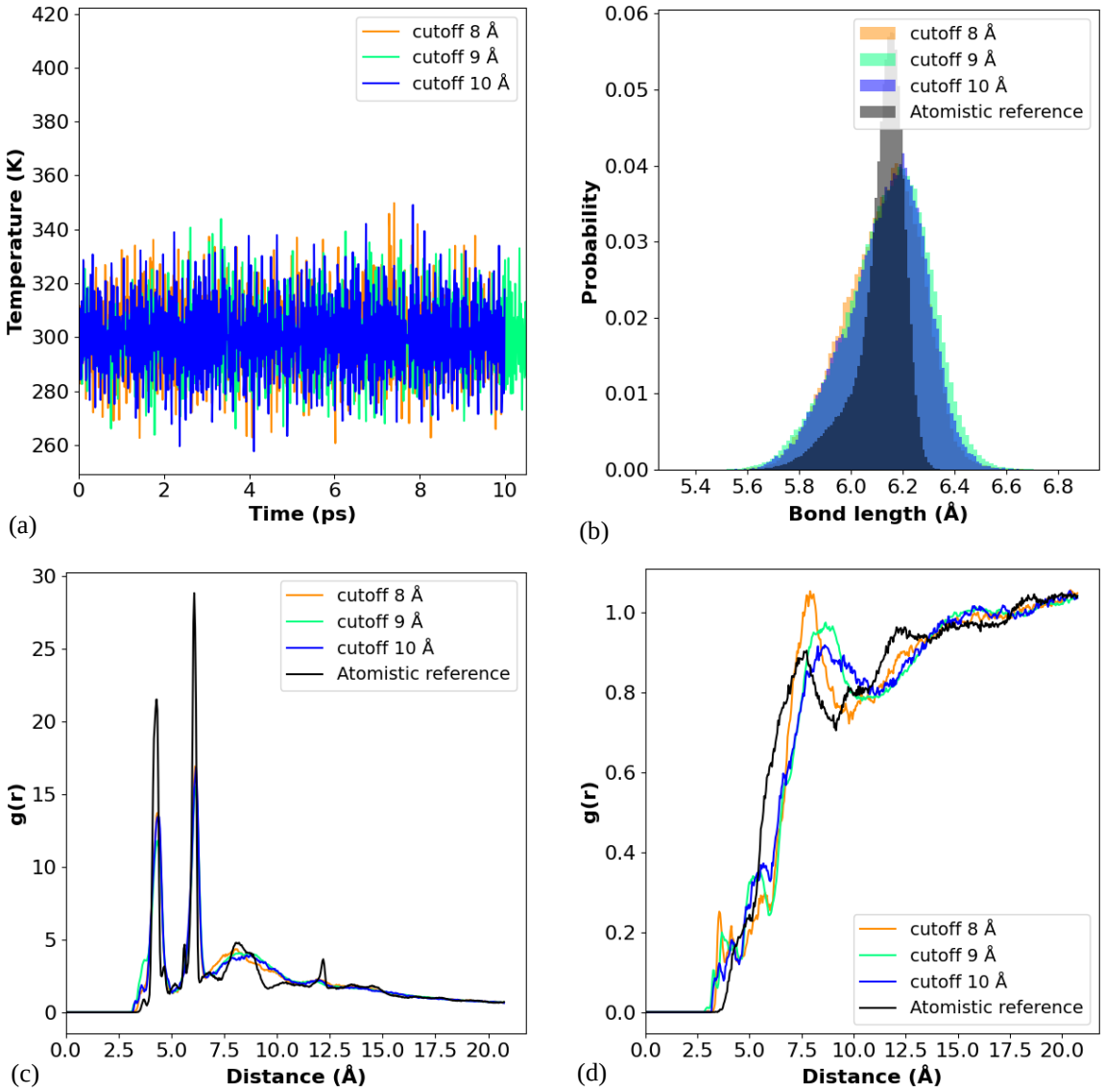

Figure 8: Comparison of the (a) temperature trend, (b) bond length probability distribution for one of the backbone bond types (bond between a "red" and a "violet" bead – see Figure 1), (c) intramolecular radial distribution function, and (d) intramolecular distribution function during CG NVT MD simulations of PIM-1 mapped using 3 CG beads per molecule, performed at 300 K and atmospheric pressure with GCNN potentials trained using Eq. 5 as the loss function, with energy scaling factor λ=0.05, and various values of the local neighbourhood cutoff radius (orange: 8 Å, green: 9 Å, blue: 10 Å, black atomistic reference).

4. Conclusions

In this work, we explored the application of graph convolutional neural networks for the generation of coarse-grained potentials for two polymer systems, namely polyethylene and PIM-1, at bulk conditions. For polyethylene, two different CG mappings were considered. We investigated the effect of the model hyperparameters on the results and identified a suitable combination for each system. In the case of polyethylene, simulations of systems containing longer chains were performed using the best model, and consistent results were obtained. The code developed and the best models obtained are available online, and links can be found in the Supplementary Materials section.

Future work will focus on the investigation of methods to lower the burden of hyperparameter optimization, for example by implementing automatic schemes for the balancing of the loss function coefficients [23], and bypassing the definition of adjustable prior energy terms. Moreover, it will be of interest to test different architectures to build the feature representation of the molecules, such as those based on an equivariant description of the local environment including distance vectors in addition to distance values [24].

This work attests to the potential of using Machine Learning within multiscale molecular simulation frameworks of complex systems, to expand the scope of applicability of molecular modelling to materials of industrial relevance and enable the creation of novel multiscale simulation pipelines. Given the early stage of research and limited available reports on the utilization of machine learning potentials in CG molecular simulations, it is crucial to uncover and discuss not only the progress but also the obstacles, challenges, and any unforeseen or problematic behaviour that may arise. Sharing such insights will facilitate the pursuit of enhanced strategies, accelerating the pace of discovery and broadening the scope of successful applications.

Supplementary Material

The code used in this work is available on GitHub at the following link: https://github.com/ml-multimem/schnetpack-for-bulk-systems. LAMMPS input data to perform the simulations that were used as training sets can be downloaded at the following link: https://zenodo.org/records/10352368. The best model obtained for each material, as well as full specifications of their hyperparameters, can be found at the following link: https://zenodo.org/records/10343599.

Acknowledgments

E.R. gratefully acknowledges funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 101030668.

This work was supported by computational time granted from the National Infrastructures for Research and Technology S.A. (GRNET S.A.) in the National HPC facility - ARIS - under the project MULTIPOLS (ID:011032), MULTIPOLS II (ID: 013019) and ML-SOFT (ID: 015020).

References

[1] J. Pfaendtner, R.Q. Snurr, V. Van Speybroeck, Editorial overview: Frontiers of chemical engineering: Molecular modeling, Curr. Opin. Chem. Eng. 23 (2019) A1–A2. https://doi.org/10.1016/j.coche.2019.06.001.

[2] D.N. Theodorou, Hierarchical modelling of polymeric materials, Chem. Eng. Sci. 62 (2007) 5697–5714. https://doi.org/10.1016/j.ces.2007.04.048.

[3] A. Gooneie, S. Schuschnigg, C. Holzer, A Review of Multiscale Computational Methods in Polymeric Materials, Polymers (Basel). 9 (2017) 16. https://doi.org/10.3390/polym9010016.

[4] S.Y. Joshi, S.A. Deshmukh, A review of advancements in coarse-grained molecular dynamics simulations, Mol. Simul. 47 (2021) 786–803. https://doi.org/10.1080/08927022.2020.1828583.

[5] E. Brini, E.A. Algaer, P. Ganguly, C. Li, F. Rodríguez-Ropero, N.F.A. van der Vegt, Systematic coarse-graining methods for soft matter simulations – a review, Soft Matter. 9 (2013) 2108–2119. https://doi.org/10.1039/C2SM27201F.

[6] E. Kocer, T.W. Ko, J. Behler, Neural Network Potentials: A Concise Overview of Methods, Annu. Rev. Phys. Chem. 73 (2022) 163–186. https://doi.org/10.1146/annurev-physchem-082720-034254.

[7] E. Ricci, N. Vergadou, Integrating Machine Learning in the Coarse-Grained Molecular Simulation of Polymers, J. Phys. Chem. B. 127 (2023) 2302–2322. https://doi.org/10.1021/acs.jpcb.2c06354.

[8] K.T. Schütt, P.J. Kindermans, H.E. Sauceda, S. Chmiela, A. Tkatchenko, K.R. Müller, SchNet: A continuous-filter convolutional neural network for modeling quantum interactions, in: I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Adv. Neural Inf. Process. Syst., Curran Associates, Inc., 2017. https://doi.org/https://proceedings.neurips.cc/paper/2017/file/303ed4c69846ab36c2904d3ba8573050-Paper.pdf.

[9] P.M. Budd, B.S. Ghanem, S. Makhseed, N.B. Mckeown, K.J. Msayib, E. Tattershall, C.E. Tattershall, Polymers of intrinsic microporosity (PIMs): robust, solution-processable, organic nanoporous materials, Chem. Commun. (2004) 230–231. https://doi.org/10.1039/b311764b.

[10] K.T. Schütt, P. Kessel, M. Gastegger, K.A. Nicoli, A. Tkatchenko, K.R. Müller, SchNetPack: A Deep Learning Toolbox for Atomistic Systems, J. Chem. Theory Comput. 15 (2019) 448–455. https://doi.org/10.1021/acs.jctc.8b00908.

[11] K.T. Schütt, S.S.P. Hessmann, N.W.A. Gebauer, J. Lederer, M. Gastegger, SchNetPack 2.0: A neural network toolbox for atomistic machine learning, J. Chem. Phys. 158 (2023) 144801. https://doi.org/10.1063/5.0138367.

[12] S. Plimpton, Fast Parallel Algorithms for Short-Range Molecular Dynamics, J. Comput. Phys. 117 (1995) 1–19. https://doi.org/10.1006/jcph.1995.1039.

[13] F. Ercolessi, J.B. Adams, Interatomic Potentials from First-Principles Calculations: The Force-Matching Method, Europhys. Lett. 26 (1994) 583–588. https://doi.org/10.1209/0295-5075/26/8/005.

[14] B.E. Husic, N.E. Charron, D. Lemm, J. Wang, A. Pérez, M. Majewski, A. Krämer, Y. Chen, S. Olsson, G. de Fabritiis, F. Noé, C. Clementi, Coarse graining molecular dynamics with graph neural networks, J. Chem. Phys. 153 (2020) 194101. https://doi.org/10.1063/5.0026133.

[15] E. Ricci, G. Giannakopoulos, V. Karkaletsis, D.N. Theodorou, N. Vergadou, Developing Machine-Learned Potentials for Coarse-Grained Molecular Simulations: Challenges and Pitfalls, in: Proc. 12th Hell. Conf. Artif. Intell., ACM,

New York, NY, USA, 2022: pp. 1–6. https://doi.org/10.1145/3549737.3549793.

[16] D.-P. Gerakinis, E. Ricci, G. Giannakopoulos, V. Karkaletsis, D.N. Theodorou, N. Vergadou, Machine Learning-Based Coarse Grained Interaction Potentials for Molecular Systems, (2024) Zenodo. https://doi.org/10.5281/zenodo.10501037.

[17] K.T. Schütt, H.E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, SchNet – A deep learning architecture for molecules and materials, J. Chem. Phys. 148 (2018) 241722. https://doi.org/10.1063/1.5019779.

[18] A. Hjorth Larsen, J. Jørgen Mortensen, J. Blomqvist, I.E. Castelli, R. Christensen, M. Dułak, J. Friis, M.N. Groves, B. Hammer, C. Hargus, E.D. Hermes, P.C. Jennings, P. Bjerre Jensen, J. Kermode, J.R. Kitchin, E. Leonhard Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. Bergmann Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K.S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, K.W. Jacobsen, The atomic simulation environment—a Python library for working with atoms, J. Phys. Condens. Matter. 29 (2017) 273002. https://doi.org/10.1088/1361-648X/aa680e.

[19] L.J. Abbott, C.M. Colina, Atomistic structure generation and gas adsorption simulations of microporous polymer networks, Macromolecules. 44 (2011) 4511–4519. https://doi.org/10.1021/ma200303p.

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

[21] X. Fu, Z. Wu, W. Wang, T. Xie, S. Keten, R. Gomez-Bombarelli, T. Jaakkola, Forces are not Enough: Benchmark and Critical Evaluation for Machine Learning Force Fields with Molecular Simulations, ArXiv. (2022). https://doi.org/10.48550/arXiv.2210.07237.

[22] S. Stocker, J. Gasteiger, F. Becker, S. Günnemann, J.T. Margraf, How robust are modern graph neural network potentials in long and hot molecular dynamics simulations?, Mach. Learn. Sci. Technol. 3 (2022) 045010. https://doi.org/10.1088/2632-2153/ac9955.

[23] S. Dellis, E. Ricci, D.-P. Gerakinis, N. Vergadou, G. Giannakopoulos, Evaluation scheme for self-adaptive methods of coefficients of loss components of multi-objective loss function, (2024) Zenodo. https://doi.org/10.5281/zenodo.10500940.

[24] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J.P. Mailoa, M. Kornbluth, N. Molinari, T.E. Smidt, B. Kozinsky, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, Nat. Commun. 13 (2022) 2453. https://doi.org/10.1038/s41467-022-29939-5.