THE ATOMISTIC RECONSTRUCTION OF COARSE-GRAINED POLYMERIC SYSTEMS VIA MACHINE LEARNING TECHNIQUES

by

JAKE OLSEN

A THESIS

Presented to the Department of Chemistry and Biochemistry
and the Robert D. Clark Honors College
in partial fulfillment of the requirements for the degree of
Bachelor of Science

June 2020

An Abstract of the Thesis of

Jake Olsen for the degree of Bachelor of Arts
in the Department of Chemistry and Biochemistry to be taken June 2020

Title: The Atomistic Reconstruction of Coarse-Grained Polymeric Systems via Machine Learning Techniques

Approved: _____*Marina Guenza, Ph.D*_____
Primary Thesis Advisor

The development of a statistically accurate backmapping procedure, coupled with an accurate coarse-graining (CG) method, is necessary as it would allow a system to freely transform between varying degrees of CG. This ability allows for the computational gain of CG with the resolution of atomistic simulations. Therefore, using state-of-the-art machine learning techniques coupled with atomistic simulation data, we have developed a backmapping procedure for CG polymeric systems. Specifically, we used a gated recurrent unit (GRU) to learn the atomistic structure within a single CG site of a polyethylene system. A categorical cross-entropy loss function was used to allow for more flexibility in the model. The model's training yielded consistent loss and validation loss demonstrating that the model did not overfit the data. Furthermore, the model was able to accurately reproduce a variety of structural quantities, such as the bond angle, bond length, dihedral angle, mean-square internal distance (MSID), end-to-end distance, and distribution about the center-of-mass.

Acknowledgements

I would like to extend my greatest thanks to Professor Marina Guenza, who gave endless amounts of support and energy to this project and who inspired my interests into computational chemistry through her enthusiasm and excitement in having me join her lab. My experiences in Professor Guenza's lab have pushed me as a scientist and have helped me develop countless skills. I would also like to thank the other members of the Guenza lab for the continual advice end help on this project and throughout all my research endeavors. And I would also like to thank Dr. Jake Searcy for help with the machine learning and programming components of my project. This project would not have been possible without the support from these people.

I would like to thank the Undergraduate Research Opportunity Program at the University of Oregon for giving me the amazing opportunity of becoming a member of the Presidential Undergraduate Research Scholars program, which provided me with funding, mentorship, and support throughout my senior year. I would also like to thank the Chemistry Department for awarding me the P-CHEM Undergraduate Fellowship, which provided me with funding and support over the summer of 2019.

Beyond academic support, I would like to thank my family and friends who have provided me with an amazing support network and who have uplifted me in my pursuit of academia. There are no words that can sufficiently describe my thanks to these people.

Table of Contents

List of Figures

Background and Introduction

In our modern age, the study of polymeric systems is becoming increasingly important. Polymeric systems include things like proteins, DNA, and synthetic plastics, all of which play crucial roles in our everyday lives. These systems are of great interest for their applications in material design. For example, it would be extremely useful to be able to understand how changes to a polymer's structure would affect the global properties of a polymeric system without having to synthesize the material. Being able to do this kind of investigation would allow for new materials to be efficiently created based on the desired global properties of the polymer. Furthermore, being able to easily study proteins and DNA on a microscopic level has great importance in the biomedical industry.

Unfortunately, when it comes down to investigating how polymeric system properties depend on molecular-level structure, experimental approaches do not always suffice, and they are economically expensive. Therefore, having time-efficient and inexpensive computational approaches to investigate polymeric systems on multiple scales, from the microscopic to the macroscopic level, is of great importance and necessity. Luckily, with the vast improvements to computers in the past several decades, investigating polymer systems from the microscopic to the macroscopic level has become more computationally feasible. Still, large systems that are of industrial interest, cannot be simulated even in supercomputers, because they require extensive computational resources. There are many different computational methods used in investigating these systems. A universal computational approach that allows for the investigation of time-dependent properties is molecular dynamics (MD) simulations.

Molecular Dynamics Simulations

To understand MD simulations, it is first necessary to understand molecular mechanical (MM) methods. MM methods neglect the quantum mechanical aspects of a system and treat molecules using classical Newtonian mechanics. This treatment corresponds to treating bonded and nonbonded interactions between atoms as springs. Like springs, the molecules have potential energy based on their distance from an equilibrium state. In this case, the potential energy between atoms of the system is purely a function of the distance between nuclei and is well approximated by force fields. Different MM force fields make different approximations to reproduce experimental quantities. This is where MD simulations come into play. MD simulations are a computational method that solves Newton's equations of motion using a given timestep to give the MM forces that are applied to the atoms within the system. Across multiple timesteps, a trajectory of positions in time is formed, giving the dynamic evolution of the system. In traditional MD simulations of polymeric systems, each polymer molecule is described at the atomistic level. However, to investigate the properties of large polymeric systems, MD simulations, when at atomistic resolution, are very computationally intensive. A way to overcome this problem is to perform MD simulations at coarse-grained (CG) resolution.

Coarse-Graining

CG is a method in which the local degrees of freedom (DOFs) of the molecular description are averaged out and the representation of the molecule is simplified to reduce the computational time. The averaging is conducted by taking a collection of consecutive monomers from the atomistic polymer chain and condensing them into a

single site located at the center-of-mass of those monomeric sites. This single site is known as a CG site, or blob. When the system, or polymer, is completely coarse-grained to a system of CG sites, it is said to be in its CG representation. If an entire polymer chain is condensed into a single site located at the polymer's center-of-mass, then it is known as a soft sphere. An illustration of the atomistic representation of a polymer and its corresponding CG representation is depicted in **Figure 1**.

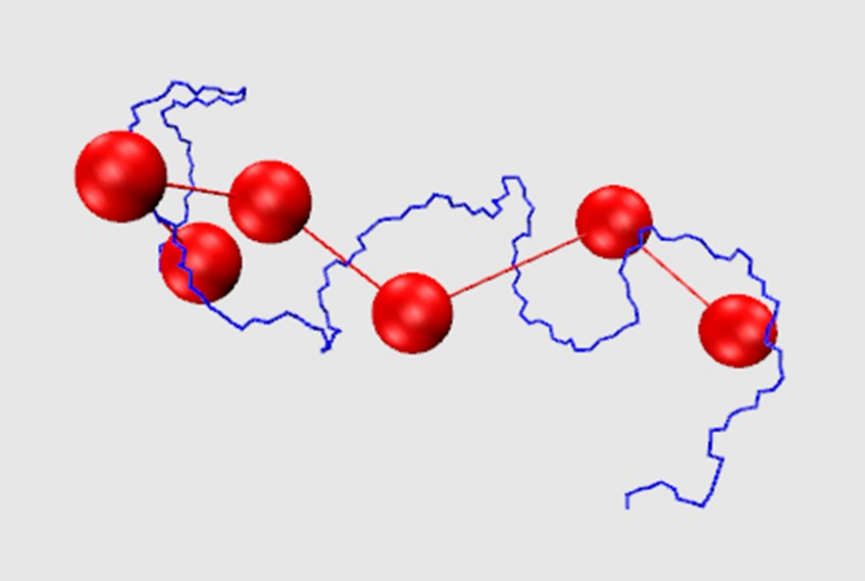

Figure 1. Atomistic and CG Representation of a Polyethylene Chain of Length 192

A polyethylene chain containing 192 monomers, depicted in blue, which has been coarse-grained to six blobs, depicted in red. Each blob contains 32 monomers of polyethylene.

The smaller number of monomers in a CG site, the finer the system is said to be, and the greater number of monomers in a CG site, the coarser the system is said to be. The process of CG is like rendering an image. The atomistic representation is like a high-resolution image while the CG representation is like a pixelated image. It is significantly quicker to render a pixelated image, but some information is lost due to pixelation. This tradeoff is the reason that CG greatly improves computational time but

loses some statistical information due to the averaging of atomic DOFs. It is important to note that in MD simulations computational time grows roughly with N^2, where N is the number of atoms in the system. Therefore, by reducing the DOFs by a factor of ten through CG methods improves computational time by a factor of 100. MD simulations can be optimized to grow as $N \ln N$ by applying a cutoff distance for the potentials and implementing a Verlet neighbor list[1,2], but CG still leads to a vast computational gain. As stated previously, some systems of industrial interest cannot even be simulated on supercomputers. A way to overcome this limit is to combine simulations of the same system but depicted at different resolutions, thus combining atomistic with CG resolution models. A depiction of different levels of CG is presented in **Figure 2**.

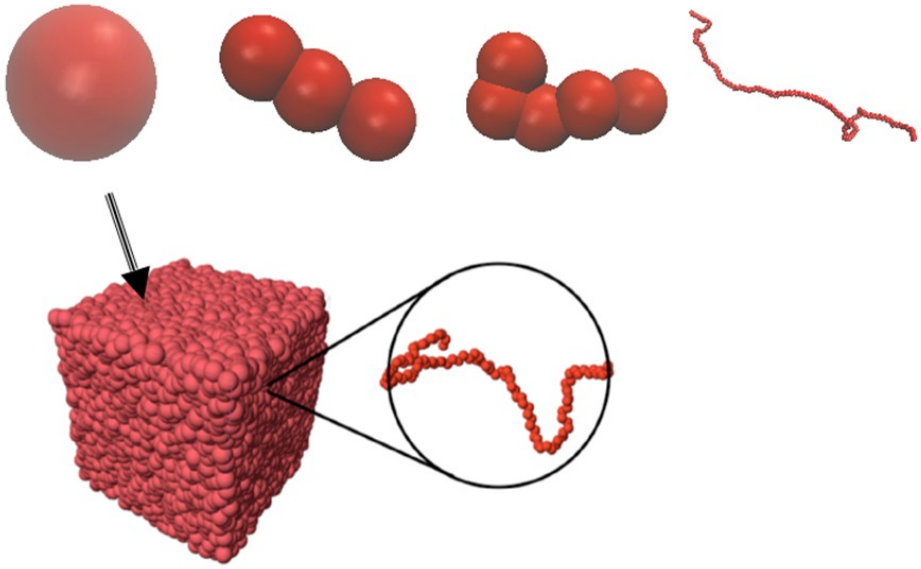

Figure 2. Polymeric System at Different Levels of CG

The far-right depicts an atomistic representation and as you go left the level of CG increases until you reach the soft sphere representation.

Once the polymer has been coarse-grained, it can be studied as it evolves in time using MD simulations. However, to conduct MD simulations, the force field, or

potential, between CG sites needs to be determined. When a system is coarse-grained, the atomistic potentials used for simulation do not apply to the CG system. Therefore, new potentials need to be derived.

There are several different coarse-graining models that attempt to determine these interaction potentials. These methods include Boltzmann inversion, iterative Boltzmann inversion, inverse Monte Carlo, and force-matching.[3–9] These methods can be classified into two broad categories: iterative and noniterative methods. Iterative methods make corrections to the CG potential by iteratively running simulations and adjusting the potential according to the expected local and global properties. Iterative Boltzmann inversion and inverse Monte Carlo are examples of iterative methods. Noniterative methods attempt to create the potential that recreates the forces on the CG sites. Boltzmann Inversion and force-matching are examples of noniterative methods. Unfortunately, these methods reproduce the structural properties of the atomistic system, but are often unable to recreate many of the thermodynamic properties within the system. Several methods require a correction to the system pressure to accommodate this issue.[3,5,6]

The main pitfall to these models is the potential is not analytically determined, which means that the calculated potential is specific to the system for which the potential has been optimized; it is not general and cannot be applied to different systems and in different thermodynamic conditions of temperature and density. On the contrary, an analytical determined CG potential is general and can be applied to a wide range of systems. Furthermore, analytically determined potentials do not require atomistic simulations to be initially performed to find the CG potential, which would largely

defeat the purpose of having a CG description in the first place. If we need to perform an initial atomistic simulation of our system, there is no purpose in performing a CG simulation as all the desired information can be obtained from the atomistic simulation.

To obtain an analytical potential for polymeric liquids, the Guenza group used liquid state theory and solved the Ornstein-Zernike (OZ) equation for the CG representation to obtain an appropriate potential for various levels of CG of the polymeric molecules. This CG approach is known as Integral Equation Coarse-Graining (IECG).

Ornstein-Zernike Equation

The total correlation function, $h(r_{12})$, is given in **Eq. (1)**.

$$h(r_{12}) = g(r_{12}) - 1 \quad \textbf{(1)}$$

$h(r_{12})$ is a measure of the influence of atom one on atom two as a function of the distance between the two atoms, r_{12}. $g(r_{12})$ is the radial distribution function, which essentially gives the radial structure of a complex isotropic system. $g(r_{12})$ can be thought of as the probability of finding a second atom distance r_{12} away from a first one. $g(r_{12})$ is zero for small distances, where the atoms cannot superimpose, and approaches one at larger distances, where the liquid is statistically uniform.

In an atomic liquid, such as liquid argon, the OZ equation decomposes $h(r_{12})$ into a direct component, $c(r_{12})$, and an indirect component. This decomposition can be seen in **Eq. (2)**.

$$h(r_{12}) = c(r_{12}) + \rho \int c(r_{13}) h(r_{32}) d\acute{r}_3 \quad \textbf{(2)}$$

ρ is the density. In **Eq. (2)**, the indirect component is the effect of particle one on particle three, which, in turn, affects particle two. This effect is then integrated over all possible positions for atom three. This decomposition is particularly useful for low-density systems, such as gas phase systems, because the integral goes to zero only leaving the direct correlation function. However, for systems in which the indirect component cannot be ignored, such as with liquids, the integral component needs to be solved. **Eq. (2)** can be rewritten as a convolution, as seen in the manipulations in **Eq. (3)**.

$$h(\acute{r}_{12}) = c(\acute{r}_{12}) + \rho \int c(\acute{r}_{12} - \acute{r}_{32}) h(\acute{r}_{32}) d\acute{r}_{32} = c(\acute{r}_{12}) + \rho (c * h)(\acute{r}_{12}) \quad \textbf{(3)}$$

Eq. (3) can be solved by applying a Fourier transform (FT). If we denote the FT of $h(\acute{r})$ and $c(\acute{r})$ as $\hat{H}(\acute{k})$ and $\hat{C}(\acute{k})$ respectively, then the FT of **Eq. (3)** can be computed by applying the convolution theorem. The result of the FT of **Eq. (3)** is given in **Eq. (4)**.

$$\hat{H}(\acute{k}) = \hat{C}(\acute{k}) + \rho \hat{H}(\acute{k}) \hat{C}(\acute{k}) \quad \textbf{(4)}$$

$\acute{k}$ is the spatial frequency. Because we are dealing with a polymeric system, **Eq. (4)** must be manipulated to include the intramolecular structure, $\hat{\Omega}(\acute{k})$, of the polymers. This inclusion was done by Schweizer and Curro in the polymer reference interaction site model (PRISM).[10] The result of including the intramolecular $\hat{\Omega}(\acute{k})$, in PRISM is depicted in **Eq. (5)**.

$$\hat{H}(\acute{k}) = \hat{\Omega}(\acute{k}) \hat{C}(\acute{k}) \left(\hat{\Omega}(\acute{k}) + \hat{H}(\acute{k}) \right) \quad \textbf{(5)}$$

Unfortunately, **Eq. (4)** and **Eq. (5)** are not in a closed form, meaning that, for example in **Eq. (5)**, $\hat{H}(\acute{k})$ cannot be written purely as a function of $\hat{C}(\acute{k})$ and $\hat{H}(\acute{k})$, because both values are unknown. Therefore, a closure relation needs to be applied to **Eq. (5)** so that it is in a usable form. Common closures include the hypernetted-chain

equation (HNC), mean spherical approximation (MSA), and the Percus-Yevick (PY) approximation. Each of these equations relate $\widehat{H}(\acute{k})$ to $\widehat{C}(\acute{k})$ and solve the OZ equation for different types of systems. Each closure equation is an approximate solution of the full equation, which holds in specific situations. For example, the MSA works well for dense liquids, the PY works well for systems that interact with sharp repulsive potentials, and the HNC works well for systems that interact with soft repulsive potential, like the IECG potentials.

Solution of the PRISM Equation and Determining the IECG Potential

The molecular OZ equation, **Eq. (5)**, was solved for small molecular fluids in the 1970's by Chandler and Anderson who expanded upon the OZ equation yielding the reference interaction site model (RISM).[10,11] Their model was designed to be applied to molecular fluids and chain clusters. In order to solve the OZ equation for polymers, RISM was further expanded to the polymer reference interaction site model (PRISM) by Schweizer and Curro.[12] PRISM was designed to be applied to a variety of polymer melts and blends. Then, Yatsenko et al.[13] applied PRISM to solve for the potential between soft spheres for the ideal model of a polymer represented as an infinite-length thread at constant liquid density. Finally, the potential for more realistic CG polymer representations was analytically solved by Clark, McCarthy, and Guenza[14,15] opening the possibility of performing long simulations of polymer liquids with a realistic CG potential, which reproduces the equation of state of the atomistic polymer liquid. The IECG model has been shown to be accurate and computationally efficient across multiple levels of CG.[4,16–19] Furthermore, it has been shown to be thermodynamically

consistent across multiple levels of CG.[16,20,21] Overall, IECG is a robust CG model that gives an accurate representation of CG systems.

However, the computational gain from coarse-graining a system is coupled with the loss of statistical information on the local length scale. The loss of statistical information is the result of the local DOFs of the atomic structure being averaged out to gain high computational efficiency. Therefore, to regain the lost atomistic information when it is required, while still maintaining computational efficiency, we envisioned a CG simulation that reinserted atomistic information where high-resolution is needed. Thus, the CG simulation trajectories need to be transformed, at an instant, during the IECG simulation to an atomistic representation, starting from the IECG configuration. This process is known as backmapping.

Backmapping

The process of backmapping is nontrivial given that an atomistic model maps onto exactly one coarse-grained model while a coarse-grained model maps onto many atomistic models. This ambiguity between coarse-grained systems and atomistic systems makes it difficult to develop a statistically accurate backmapping procedure while still maintaining the gain in computational time. If a statistically accurate method were developed, it would allow for MD simulations to transfer between atomistic and CG systems freely, improving computational efficiency while maintaining local length scale statistics. Geometric and structural approaches are ideal for backmapping because they tend to require far less computational time than optimization processes. Several backmapping procedures have been developed to accomplish this goal.[6,22–24] However, these backmapping approaches fail in the same capacity as many of the coarse-graining

models. Several of the models backmap from too fine-grained of a CG system[22,23] while the others lead to local length scale deformation, which then require further simulation to regain the correct statistics.[6,24,25] Overall, these methods are unable to recreate the expected thermodynamic, local length, and large length scale statistics of the system, which prompts the need for a statistically accurate backmapping procedure.

Therefore, my project is oriented around attempting to use state-of-the-art machine learning techniques to develop a computationally efficient backmapping procedure to reconstruct the atomistic information from the CG trajectory. This method would allow for MD simulations to transfer between atomistic and CG systems freely, improving computational efficiency.

Methods

The System

The system used in the development of my backmapping procedure is composed of 350 polyethylene chains of length 192. The system was coarse-grained so that each chain was composed of 6 CG sites each containing 32 monomer units. Polyethylene is the most basic polymeric system as it is a long chain of single-bonded, non-branching carbon atoms. The monomer unit of polyethylene is a single methyl group, which is composed of one carbon atom and two hydrogen atoms, or three hydrogen atoms if at the end of the polymeric chain.

I performed an atomistic MD simulation of polyethylene using LAMMPS on the Comet Supercomputer at the San Diego Supercomputing Center. The atomistic simulation is performed to test and assess the accuracy of the backmapping procedure. It was also used to produce the CG representation used to train the machine learning model. The system underwent 1 ps of simulation using a soft potential to remove the extremely unrealistic structures. Then, the system underwent a short 25 ps equilibration minimization using a Leonard-Jones (LJ) potential with a cutoff distance of 14.0 Å. After minimization, 160 ns of production, using the same LJ potential, was run starting from the final configuration of the equilibration. All simulations were conducted in the NVT ensemble, with the temperature controlled by the Nosé-Hoover thermostat.[26,27]

Geometric Approach

Prior to the development of the machine learning model presented here, I attempted to develop a geometric approach for a backmapping procedure. The

geometric backmapping procedure was developed in Python 3. The backmapping procedure works by first removing one CG site from the initial CG system. Then, one atomistic site is inserted between each pair of CG sites and at both ends of the CG chain. This process functionally doubles the number of sites in the chain. This process is iteratively repeated until the desired atomistic resolution is achieved. However, because one site was removed at the beginning of this process, one site needs to be added at the end of the backmapping procedure to get to the desired resolution. Because this process doubles the number of sites at each iteration, it is necessary that the number of target atomistic sites within a polymer is a multiple of two. Ideally, it would be at least a multiple of 2^5 so that there is a minimum of 30 monomers per CG site. This minimum number is the result of assuming that $\Omega(\acute{r})$ is Gaussian to analytically solve for the IECG potential.

We named this process by which sites are inserted, the Iterative Cone Method (ICM), which was developed by the Guenza group. A depiction of how this method functions is seen in **Figure 3**.

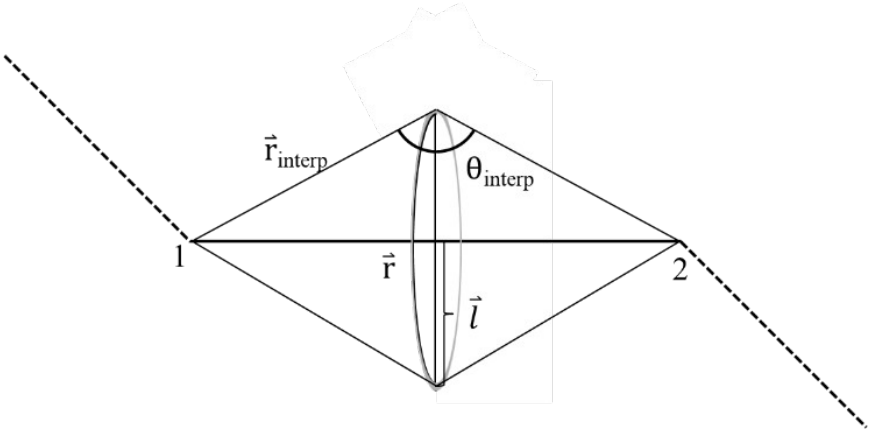

Figure 3. Diagram of the ICM Process

$\acute{r}$ is the bond length between CG sites 1 and 2, $\acute{r}_{interp}$ is the interpolation bond length, θ_{interp} is the interpolation bond angle, and $\acute{l}$ is the interpolation radius determined from $\acute{r}_{interp}$ and θ_{interp}.

The ICM works by using equilibrated atomistic data to determine the bond length and bond angle as a function of CG. Then, $\acute{r}_{interp}$ and θ_{interp} are chosen so that when the site is inserted, it will have statistically consistent bond length and bond angle with respect to that iteration. Then, $\acute{l}$ is determined as a weighted average of $\acute{r}_{interp}$ and θ_{interp}. The weighted average is conducted based on the given iteration. Finally, a site is inserted directly in between sites 1 and 2, somewhere on the cone with base radius, $\acute{l}$.

An algorithmic flowchart depicting the main components of the backmapping procedure is depicted in **Figure 4**.

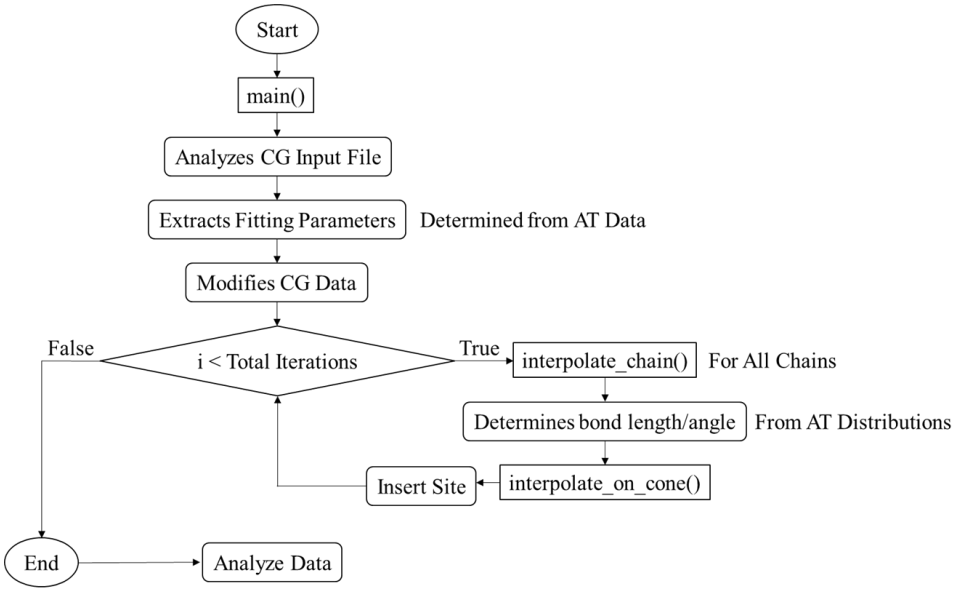

Figure 4. Algorithmic Flowchart for the ICM Backmapping Procedure

The ovals represent the beginning and end of the main procedure. The rectangles with sharp edges represent the main called functions in the procedure. The rectangles with rounded edges represent general processes that occur in the procedure. The diamond represents a for loop.

As seen in **Figure 4**, the procedure starts by reading in the CG data and then extracting some various fitting parameters that were determined from the equilibrated atomistic data. Then, the chain is modified by doing an interpolation where sites are only added between adjacent CG sites. Then, the original CG sites are deleted. This process reduces the total number of CG sites by one, but it means that the modified chain sites correspond to actual atomistic sites rather than CG sites, which correspond to the center-of-mass of several sites. Also, reducing the number of CG sites by one allows for sites to be added at both ends of the chain throughout the iterative process resulting in a more symmetric construction process. Once the chain is modified, the iterative process can begin. For each iteration, sites are inserted using interpolate_on_cone(), which is repeated for each pair of sites and for each polymer chain, which is the purpose

of interpolate_chain(). After the desired resolution is reached, the procedure ends, and the data can be analyzed.

This code was made available as a free source code repository on the GitHub platform. The repository link is https://github.com/jake93936/Backmapping_ICM.

This approach underwent considerable development, but it was unable to fully recreate the desired statistics and it had varying limitations. The problem with this approach is that the insertion of a new atom is local but has long-range effects on the distribution of the monomers in the chain. In fact, the insertion of one monomer can affect the position of another monomer in a distant location along the chain, which can be in close spatial proximity due to the formation of loops in the chain. This combination of local and long-range effects renders the problem of optimization very complex to solve. In practice, this geometrical approach can only work for short chains. Therefore, a machine learning approach was developed to solve this complex optimization problem.

Machine Learning Approach

Machine learning is an application of artificial intelligence (AI) that allows a model to automatically learn and improve from experience. Machine learning is a useful approach because the user is not required to manually code what the model needs to learn. Instead, the user sets certain parameters and the model learns what information is important in a system. Machine learning methods accomplish this task by using an artificial neural network (ANN), which functions like a biological neural network or brain. A basic diagram of a machine learning neural network is given in **Figure 5**.

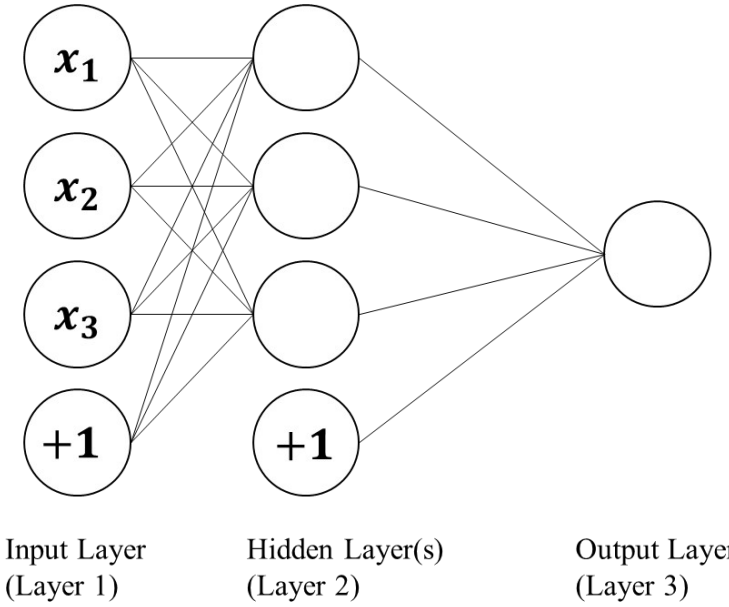

Figure 5. Diagram of an Artificial Neural Network

Each circle corresponds to a node within the ANN. The first layer is the input layer, which contains all the information fed into the model. The hidden layers determine what information is important from the first layer. The output layer is the predicted output from the network.

An ANN is composed of sets of nodes, depicted as circles in **Figure 5**, which act as individual neurons that turn on or off depending on what information they receive from the previous layers. The received information is represented by the connecting lines in **Figure 5**. The ANN takes in a set of data, represented as x_1 through x_3 in **Figure 5**, which get fed forward through the hidden layers. The hidden layers help determine what information is important from the inputs and then passes on that information in the form of an output. The model learns by adjusting the weights applied to the connections between all the nodes until the network reliably produces the correct output. The +1 values in the nodes in **Figure 5** are bias terms which give the model more flexibility in the outputs it can predict.

The model used for my project is a variant of a Recurrent Neural Network (RNN). RRNs are a class of neural networks that conserve “memory”. A standard RNN suffers from something known as the vanishing gradient problem. This problem occurs when determining how to update the weights of the nodes in the hidden layers. During back propagation, a process by which the weights are updated, the gradient can decrease to zero or explode to infinity which is a problem in updating the node weights. Therefore, a gated recurrent unit (GRU) network was implemented. A GRU solves the vanishing gradient problem by having something known as an update gate and a forget gate. These gates more reliably determine what information to keep and what information to forget from the past outputs of the model.

The initial goal with the machine learning approach was to backmap a single CG site rather than an entire chain to test the feasibility of this approach. The idea was that the model would predict atomistic sites one at a time within the CG site until the entire atomistic structure was reinserted. A GRU was chosen for this model because the model would be able to store a form of “memory” from the past site insertions. This memory would better inform the model on how to place the following sites. The specific model used is an encoder-decoder sequence to sequence GRU model. This specific model takes in a sequence of data and outputs a sequence of data that may be of a different size. i.e. the model predicts a single site given a set of input data that is of a different size.

The input data is configured so that the CG site that is being backmapped is placed at (0, 0, 0) on a cartesian grid. For simplicity let us call this CG site C_0. The network will receive from the atomistic simulation, as input information, the ensemble

of the positions of the actual atomistic sites within one CG site, the location of all the CG sites on the same chain as C_0, and all the CG sites not on the same chain that are within 10 Å from C_0. This information provides the model with the statistical probability of finding an atomistic site relative to the position of the CG site it belongs to, and the relative position of intramolecular and intermolecular blobs around the CG site. The model also receives the previously predicted atomistic sites within that blob, which are related to the site that we want to insert by some statistical distribution of bond lengths, bond angles, and dihedral angles. The output of the network is composed of all the predicted atomistic sites within C_0. The sites are predicted one at a time, allowing for the GRU to maintain a memory of the previously predicted sites. The model predicts probability distributions of where the site is located. We chose the model to predict probability distributions rather than discrete values, because there are many possibilities for an atomistic chain within a single CG site meaning there are many locations for each atomistic site to be placed. Because the model is predicting a distribution, the loss function being used is a categorical cross-entropy loss function. This loss function was chosen because it compares probability distributions of the predicted output of the model to the desired output. Therefore, the network is optimized so that it predicts the correct distribution of outputs rather than single discrete values. The specific output can then be chosen from that distribution appropriately. The code for the machine learning approach was written in Python 3 using TensorFlow and Keras.

Data Representation

The atomistic sites were represented in two different coordinate systems. Changing the coordinate system can help reduce the size of the data that is fed to the model, which, in turn, can make it easier for the model to learn. Therefore, we chose to investigate the performance of two coordinates systems. The first system was a spherical coordinate system where the radius was assumed to be fixed at 1.54 Å. This length corresponds to the equilibrium bond length of polyethylene. θ and ϕ were measured with respect to the previous site on the chain meaning that this is a relative coordinates system where the center of the grid is the previous site. The other coordinate system was a cartesian coordinate system where the *x*, *y*, and *z* coordinates were measured with respect to the previous site on the chain. When these coordinate systems are implemented into the model the resulting models are referred to as the spherical model and the cartesian model, respectively. Both models received and predicted coordinates in these coordinate systems. The inputs and outputs for both coordinate systems were binned to create a probability distribution over the possible input and output values. Binning allows for the use of a categorical cross-entropy loss function. As a result, the model predicted the distributions of these variables rather than a single value. Both models were trained for a total of 50 epochs. Each epoch corresponds to the model training on the entire set of training data. The entire set of data was composed of 3,200 timesteps sampled from the 160 ns of atomistic data. For each timestep, 100 individual CG sites were used. Therefore, the entire data set corresponded to 320,000 single CG examples. The training data set is composed of 80% of the total data set. The remaining 20% is reserved for development and testing.

Results and Discussion

Ideally, we would like our model to reproduce the expected thermodynamic, local length, and large length scale statistics of the system because these quantities are not accurately predicted by other backmapping procedures.[6,24,25] The thermodynamic properties are related to the efficacy of the CG method as it is the CG system that undergoes MD equilibration. Also, the backmapped system maps directly onto the original CG system. As discussed previously, IECG captures the thermodynamic properties of the system. Thus, the backmapping process should not interfere with these quantities. Therefore, we just need to investigate the models' ability to recreate local and large length scale statistics.

Model Efficacy

Prior to considering the local and large length scale statistics, it is necessary to check that the model accuracy was improving with increased learning. Also, it is necessary to see if the model is overfitting the data. Overfitting is the process in which a model begins to memorize the statistics of the given data set rather than learning general behaviors. Overfitting can be seen when a model performs better on the training than on the validation/test data set. To investigate these aspects of the model, the loss and the validation loss were plotted as a function of the epoch for the cartesian and spherical models, which can be seen in **Figure 6** and **Figure 7**, respectively.

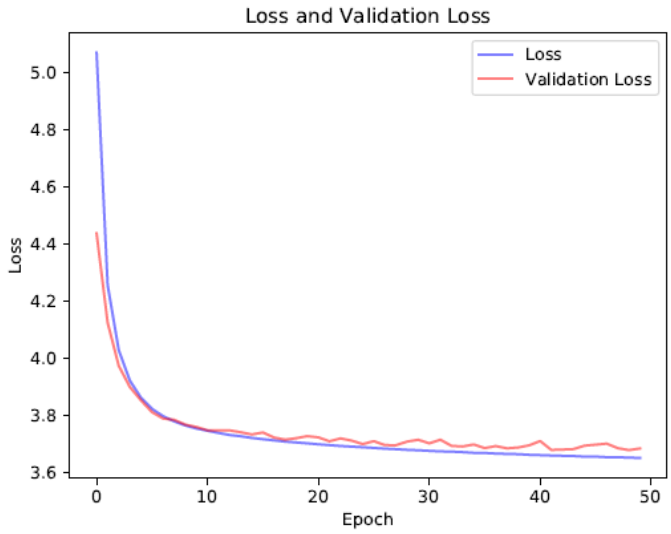

Figure 6. Loss and Validation Loss for the Cartesian Model

Loss and validation loss for the cartesian model over the 50 epochs of training.

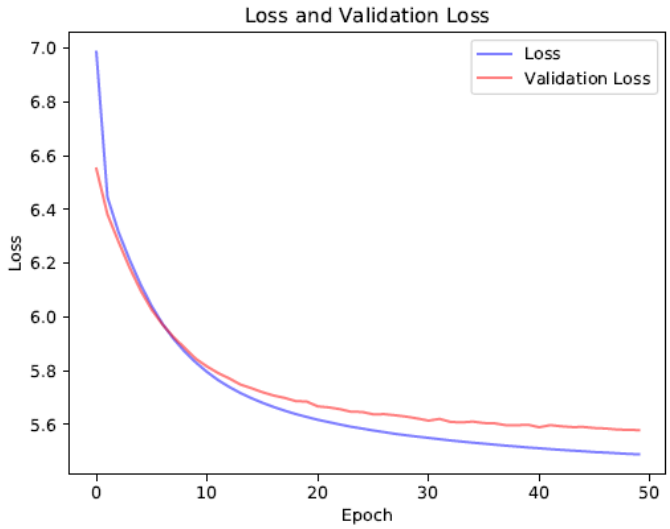

Figure 7. Loss and Validation Loss for the Spherical Model

Loss and validation loss for the spherical model over the 50 epochs of training.

As seen in **Figure 6**, the cartesian model shows strong agreement between the loss and the validation loss demonstrating that the model is not overfitting the data. Also, the loss seems to approach a minimum implying that the model will see negligible

improvement from further training. The spherical model loss and validation loss, depicted in **Figure 7**, exhibits similar behavior, but the loss is still minorly sloping downwards implying there might be more improvement to the model from more training. Also, the spherical model's validation loss does not overlap the loss as the number of epochs increases. It is important that the validation loss follows the same general trend as the loss, but it is important to take note of the discrepancy between the two, which could imply overfitting.

Local Length Scale Statistics

Now that the model efficacy has been investigated, we can consider the local length scale statistics. These quantities include the bond length, bond angle, and dihedral angle. The bond length distributions for the cartesian and spherical models are depicted in **Figure 8** and **Figure 9**, respectively. To determine the bins for all the histograms in this analysis, the Freedman Diaconis Estimator (FDE) was used on the predicted data set to determine the appropriate number of bins. This estimator works well for determining the number of bins for large data sets.

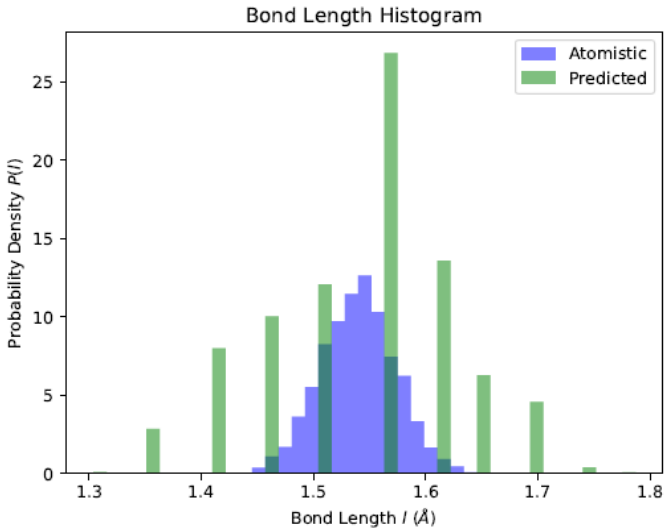

Figure 8. Bond Length Distribution for the Cartesian Model

The bins were determined using the FDE on the predicted data set. The atomistic data set was then plotted using the same bins.

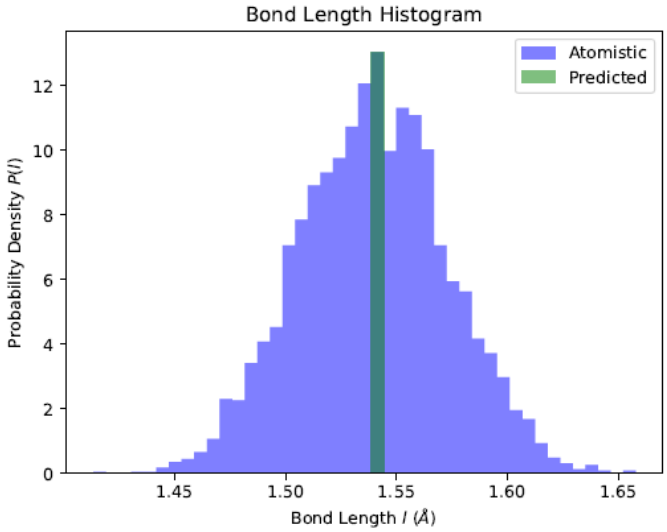

Figure 9. Bond Length Distribution for the Spherical Model

The bins were determined using the FDE on the atomistic data set. The predicted data set was then plotted using the same bins. Since that bond length was fixed for the spherical model, the predicted bond lengths only correspond to a single value. Thus, the predicted bond lengths are depicted as a single bin with the height of the corresponding atomistic bin.

As can be seen in **Figure 8** and **Figure 9** the predicted bond lengths do not recreate the atomistic bond lengths. The cartesian model produced a limited number of discrete values but appeared to have the correct mean. The spherical model produced a single bond length corresponding to polyethylene's equilibrium bond length. However, it is not an issue that the model does not perfectly reproduce the bond length statistics. Bond lengths equilibrate on the femtosecond timescale, so a short MD simulation could be run after backmapping to adjust the bond lengths.

The bond angle distributions for the cartesian and spherical model are depicted in **Figure 10** and **Figure 11**, respectively.

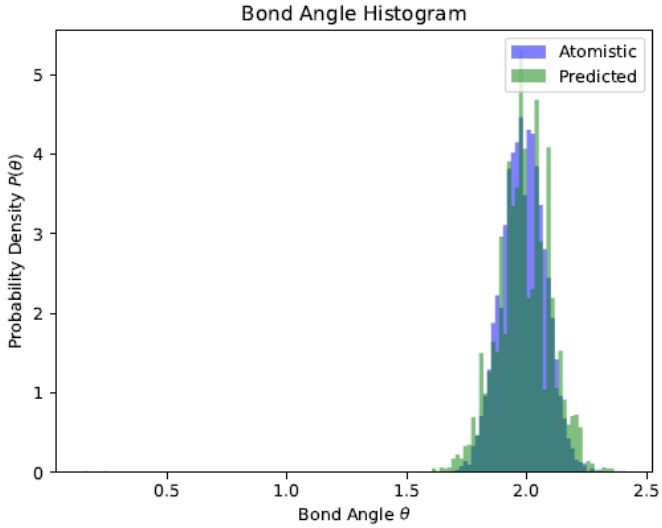

Figure 10. Bond Angle Distribution for the Cartesian Model

The bins were determined using the FDE on the predicted data set. The atomistic data set was then plotted using the same bins.

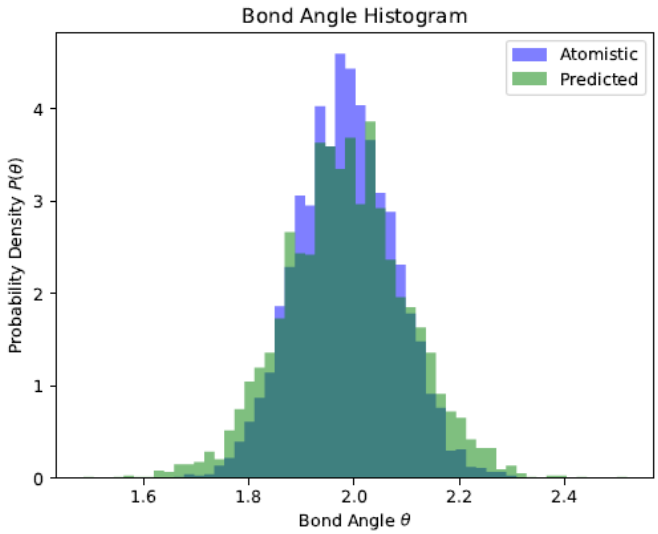

Figure 11. Bond Angle Distribution for the Spherical Model

The bins were determined using the FDE on the predicted data set. The atomistic data set was then plotted using the same bins.

As seen in **Figure 10** and **Figure 11**, the predicted bond angle distributions follow the same general trend as the atomistic distribution. The cartesian model predicted unrealistic bond angles between 0 and $\pi/2$radians. As a result, the distribution in **Figure 10** is not centered. It is unclear what is causing these unrealistic bond angles, but they do quantify a minuscule proportion of the total bond angles. Also, like the case with bond lengths, a short MD simulation could fix any minor issues with the bond angles. A short MD simulation would also help to correct the broadening of the predicted bond angles for the spherical model, as seen in **Figure 11**.

The dihedral angle distributions for the cartesian and spherical model are depicted in **Figure 12** and **Figure 13**, respectively.

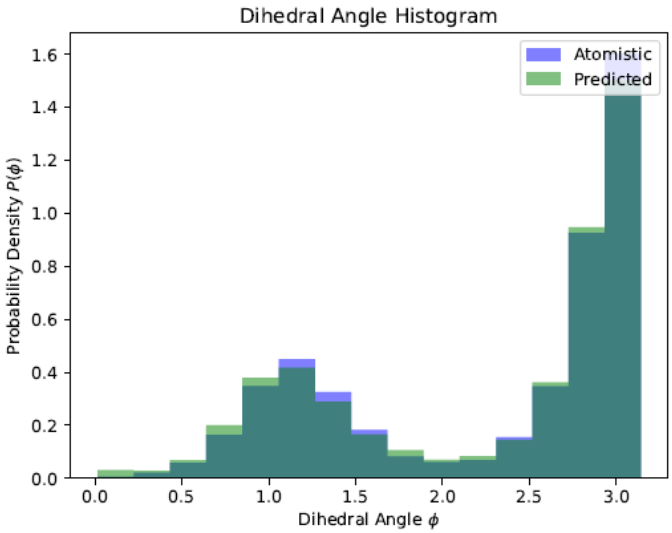

Figure 12. Dihedral Angle Distribution for the Cartesian Model

The bins were determined using the FDE on the predicted data set. The atomistic data set was then plotted using the same bins.

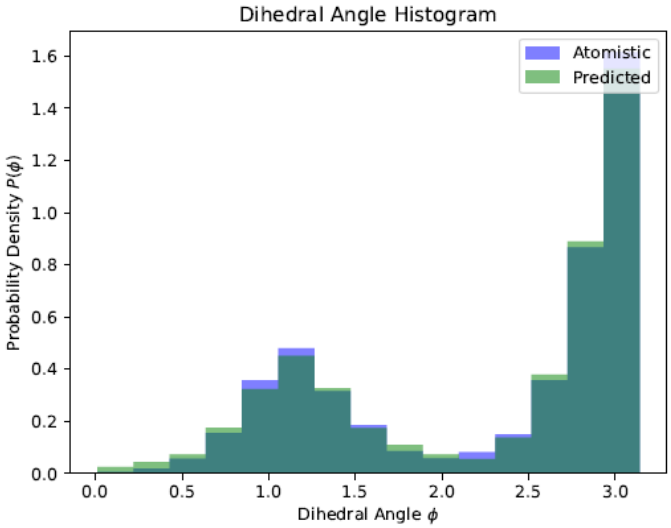

Figure 13. Dihedral Angle Distribution for the Spherical Model

The bins were determined using the FDE on the predicted data set. The atomistic data set was then plotted using the same bins.

Both models produced highly consistent dihedral angle distributions when compared to the atomistic distributions. Any minor differences can be corrected from a

short MD simulation. However, in a simulation you will see minor variation in these quantities over time, so these predicted values are accurate. Overall, both models can accurately predict the dihedral angle distributions.

The cartesian and spherical models are both able to generally recover the local length scale statistics. Any minor variations can be corrected from a short MD simulation. However, it is necessary to consider any discrepancies in determining the model's capabilities. For example, the cartesian model produced highly discretized bond length values because of the binning process of the x, y, and z coordinates. This issue gives insight into any disadvantages to this model.

Large Length Scale Statistics

The models' ability to reproduce the large length scale statistics is of great importance because large length scale statistics define the global properties of the system. Furthermore, the ICM method was unable to reproduce many large length scale statistics. To start, the end-to-end distance of a polymer chain gives a general picture of the size of a polymer. The end-to-end distance distributions for the cartesian and spherical model are depicted in **Figure 14** and **Figure 15**, respectively.

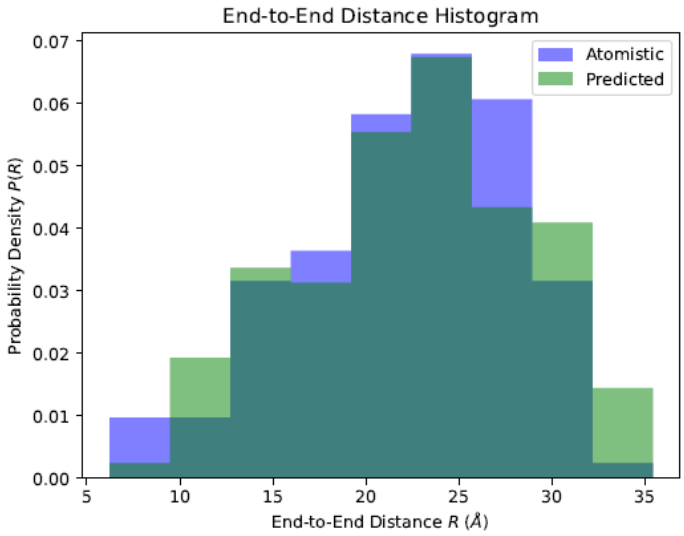

Figure 14. End-to-End Distance Distribution for the Cartesian Model

The bins were determined using the FDE on the predicted data set. The atomistic data set was then plotted using the same bins.

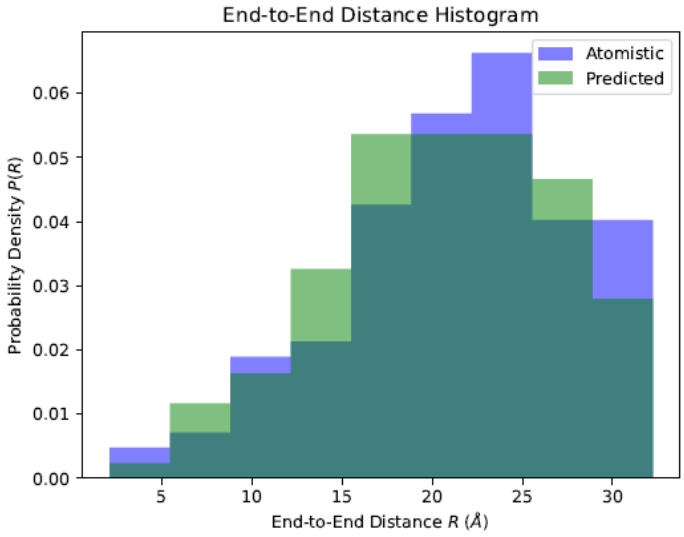

Figure 15. End-to-End Distance Distribution for the Spherical Model

The bins were determined using the FDE on the predicted data set. The atomistic data set was then plotted using the same bins.

Since each backmapping example gives exactly one end-to-end distance measurement, the distribution for end-to-end distance tends to see more variation than

other quantities. **Figure 14** and **Figure 15** show reasonable agreement between predicted and atomistic distributions for both models. Since end-to-end distance distributions tend to vary more significantly, the general relationship seen in **Figure 14** and **Figure 15** demonstrate that both models can reproduce this quantity.

The distance around the blob center is an important quantity as it gives insight into the internal structure of a CG site. The distribution of the distances around the blob center should be roughly Gaussian. However, CG sites at the end of the chain are only fixed on one end to the chain, meaning the chain is more able to stretch out and get further from the blob center. This asymmetry can be seen in **Figure 16** and **Figure 17**.

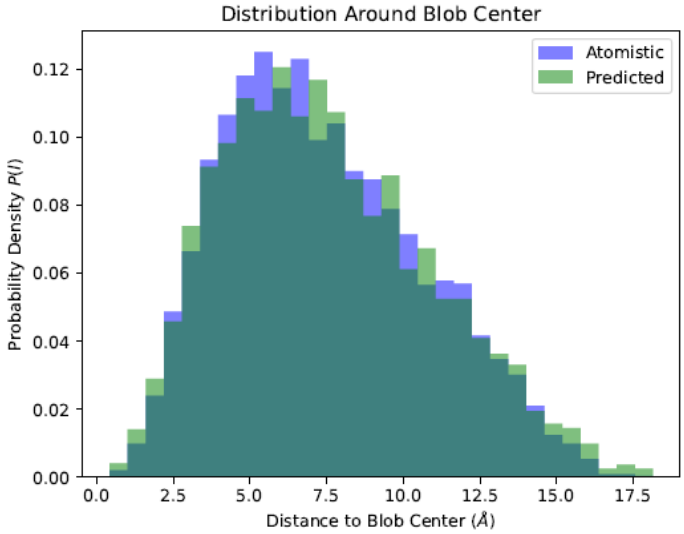

Figure 16. Distribution Around the Blob Center for the Cartesian Model

The bins were determined using the FDE on the predicted data set. The atomistic data set was then plotted using the same bins.

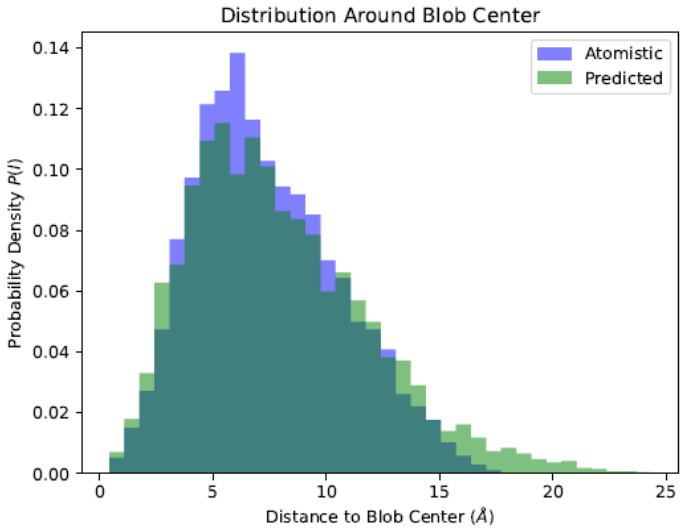

Figure 17. Distribution Around the Blob Center for the Spherical Model

The bins were determined using the FDE on the predicted data set. The atomistic data set was then plotted using the same bins.

Both models show strong agreement with the atomistic distribution around the blob center. The spherical model, as seen in **Figure 17**, does predict too large of distances. This issue gives insight into a potential downfall of the spherical model's ability to predict larger length scale properties.

Another important large length scale statistic is the mean-square internal distance (MSID). The MSID gives the average squared distance between two sites divided by the number of bond lengths between the sites. This quantity is averaged over the entire chain. When the number of bond lengths between sites is 1, then the MSID can be related to the bond length. The same is true for end-to-end distance when the number of bond lengths is the number of sites in a chain. An idealistic model, known as the freely rotating chain (FRC), assumes a fixed bond length and bond angle but allows for rotations about the bond axis. The FRC model gives a baseline to compare both the

predicted and atomistic MSIDs against. For this specific system, the FRC is an upper bound to the atomistic MSID meaning the FRC always predicts a higher value. The plots of the MSID for the cartesian and spherical models are depicted in **Figure 18** and **Figure 19**, respectively.

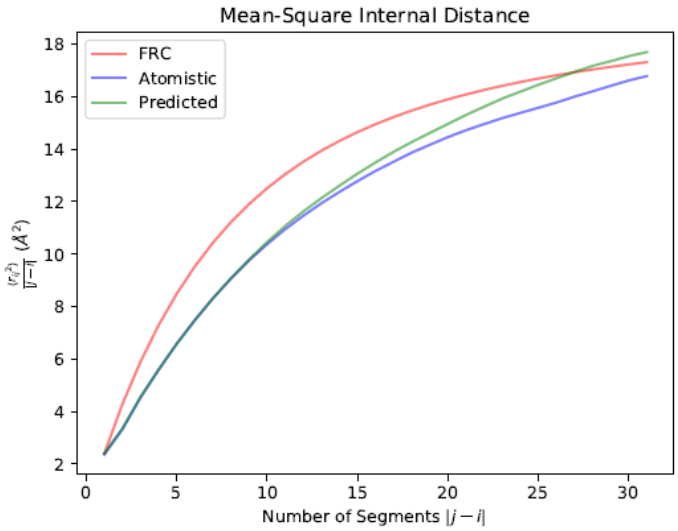

Figure 18. Mean-Square Internal Distance for the Cartesian Model

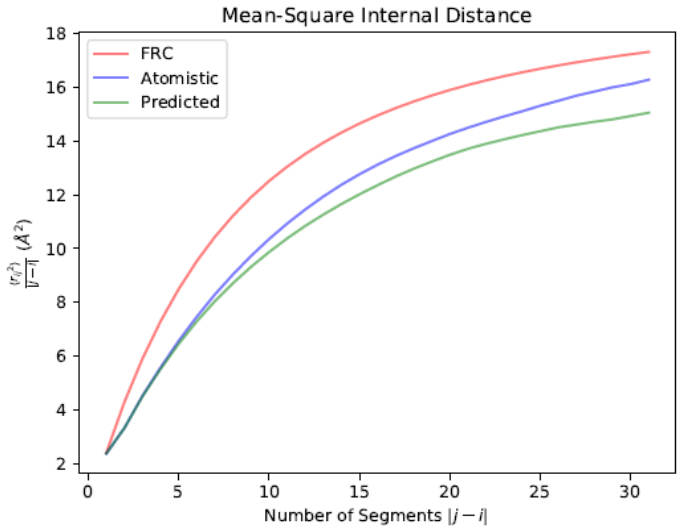

Figure 19. Mean-Square Internal Distance for the Spherical Model

The cartesian model's MSID predictions, as seen in **Figure 18**, consistently overapproximated the atomistic MSID. The spherical model's prediction, **Figure 19**, did the opposite. However, upon repeated tests of both models, the models can more accurately predict the MSID than what is depicted in **Figure 18** and **Figure 19**. This issue could be a limitation of training. i.e. the model was able to learn and recreate local length scale statistics but not long length scale statistics. Further investigation is needed to determine the cause of this effect and to determine if it is even an issue.

Both models were generally able to reproduce larger length scale statistics. However, they both displayed some issues, mainly in the prediction of the bond length and large-scale chain statistics. Also, it is important to note that the models are only recreating a single CG site rather than a chain of CG sites meaning that these quantities do not incapsulate the overall structure of the polymer system. Including the calculations for multiple CG sites would be the next step in this research project once

the minor discrepancies in bond length distribution and long-chain statistics are further minimized.

To further convey the performance of the models, the results from the machine learning backmapping procedure is depicted in **Figure 20**.

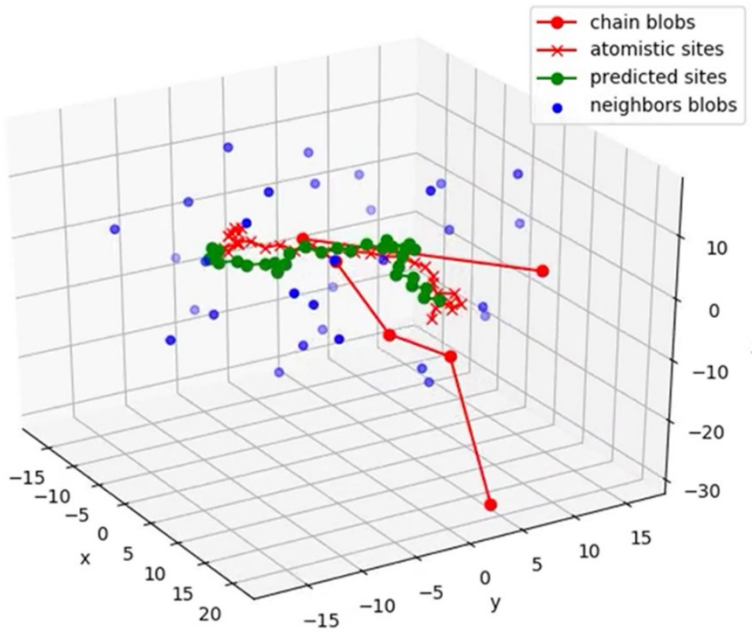

Figure 20. Visualization of a Backmapped CG Site

The red dots convey the chin blobs. The predicted sites are depicted as green dots. The atomistic sites are shown as red x's. the blobs not on the same chain are shown as blue dots and are shaded to convey depth.

The chain of CG sites is depicted in red with the position of the CG sites described as red dots. The atomistic chain that is reconstructed by the machine learning procedure is around the central CG site, C_0. The real chain in the original atomistic simulation, with it center at C_0, is depicted in red with red crosses for the atom positions. The atomistic chain predicted by our backmapping procedure is depicted in green with green dots for the position of the atoms in the atomistic description. The position of the CG sites belonging to other chains, surrounding the chain that is

reconstructed, are in blue. The predicted chain shows a reasonable structure, with realistic bond lengths, angles, and dihedral angles. We notice that the predicted chain is not directly matching the original atomistic chain; however, this is expected as we know that many possible atomistic chains correspond to the CG site C_0, and we are predicting just one of them. Overall, this result is quite promising, as it shows that the designed machine learning procedure is useful in predicting realistic atomistic chains from the knowledge of the position of the CG sites in the simulation, using just the atomistic chain statistics as inputs.

Conclusion

Overall, both models were able to give a reasonable reconstruction of the local and long length scale statistics for the system. Both models did exhibit some minor issues, but further investigation may be needed to determine if they are inherent to the models or if extended training of the networks will fix the problem. In future, the model will be extended to rebuild entire chains of blobs, which will allow the model to backmap the entire polymeric system. Being able to rebuild the entire system would allow for systems to transfer freely from the atomistic to their CG representation and back, while maintaining the thermodynamics and structural properties. This ability would further improve the power of MD simulations. Furthermore, this process would allow for multi-resolution simulations to be conducted, further improving the power of MD simulations. These abilities would also extend the versatility of IECG theory.

Glossary

Artificial Intelligence (AI): Is a branch of computer science that is concerned with developing machines that can mimic and reproduce certain task that usually require human intelligence.

Bond Angle: The angle formed between three adjacent sites, whether they are atoms, monomers, or CG sites.

Bond Length: The distance between two adjacent sites, whether they are atoms, monomers, or CG sites.

Categorical Cross-Entropy: Is a measure of the similarity between two probability distributions that correspond to single label categorization. This form of categorization is applicable when only one category applies to each data point.

Convolution: A convolution of two functions gives a third function that describes how one function changes with respect to the other.

Convolution Theorem: States that under certain conditions the Fourier transform of a convolution of two functions is the pointwise product of their Fourier transforms.

Degrees of Freedom (DOFs): The number of independent ways by which a dynamic system can move without violating any constraint imposed on it.

Dihedral Angle: Is a relationship between four adjacent sites, whether they are atoms, monomers, or CG sites. The relationship is the angle between two planes, where the first plane is defined by sites 1, 2, and 3 and the second plane is defined by sites 2, 3, and 4.

End-to-End Distance: The distance between the two terminal sites in a chain.

Ensemble: Collection of all the possible configurations of a given state of a system.

Equation of State: A thermodynamic equation relating the state variables that describe the system under a set of conditions.

Molecular Fluids: A molecular fluid refers to any kind of flowing or deforming system of colloidal or aggregate structures. This includes molecules in the gas and liquid phase.

Force Field: Is the functional form and parameters sets used to calculate the potential energy of a system.

Fourier Transform (FT): Decomposes a function of time (or space) into its principle temporal frequencies (or spatial frequencies).

Gaussian Distribution: Often called a bell curve or normal distribution. A Gaussian distribution is just a probability distribution of a random variable.

Gradient: Simplistically, a gradient is the direction and magnitude of the steepest increase at a given point.

Isotropic: Refers to a system with radial, angular, and azimuthal symmetry. It is essentially symmetric in all directions. In the context of a distribution function it is referring to radial symmetry as it is purely a function of radial distance.

Intramolecular: Corresponds to the interactions within a single molecule.

Intermolecular: Corresponds to the interactions between molecules.

Loss Function: Is a method for evaluating how well specific algorithms or models learn the given data and produce the desired output.

Mean-Square Internal Distance (MSID): Is the averaged square distance between two sites as a function of the number of sites separating those two sites. This quantity can be thought of as describing the internal structure of a chain.

Monomer: A single repeating unit that, when continuously added together, make up a polymer.

Newtonian (Classical) Mechanics: Is the branch of mechanics that is completely derived from Newton's equations of motion. i.e. position and momentum.

NVT Ensemble: Is the set of parameters that represent the possible states of a system in equilibrium with constant temperature, volume, and number of molecules.

Polymer: A large molecule that is composed of a single repeating unit, known as a monomer. Types of polymers include homopolymers, which are composed of a single repeating monomer; and copolymers, which are composed of two repeating units alternating in any order.

Potential: See definition for "Force Field".

Potential Energy: Is the energy stored within an atom or molecule. In the case of an atom, the potential energy would correspond to the various interactions that it has with the nearby atoms, such as bond length, bond angle, etc.

Spatial Frequency: Is the measure of the frequency of oscillation over a given unit of distance.

Quantum Mechanics: Is the branch of mechanics that investigates and describes the world at an atomic level. Quantum mechanics is based around the solutions to the Schrödinger Equation.

Bibliography

(1) Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review* **1967**, *159* (1), 98–103. https://doi.org/10.1103/PhysRev.159.98.

(2) Chialvo, A. A.; Debenedetti, P. G. On the Use of the Verlet Neighbor List in Molecular Dynamics. *Computer Physics Communications* **1990**, *60* (2), 215–224. https://doi.org/10.1016/0010-4655(90)90007-N.

(3) Poblete, S.; Praprotnik, M.; Kremer, K.; Delle Site, L. Coupling Different Levels of Resolution in Molecular Simulations. *Journal of Chemical Physics* **2010**, *132* (11), 114101. https://doi.org/10.1063/1.3357982.

(4) Ohkuma, T.; Kremer, K. Comparison of Two Coarse-Grained Models of Cis-Polyisoprene with and without Pressure Correction. *Polymer* **2017**, *130*, 88–101. https://doi.org/10.1016/j.polymer.2017.09.062.

(5) Auhl, R.; Everaers, R.; Grest, G. S.; Kremer, K.; Plimpton, S. J. Equilibration of Long Chain Polymer Melts in Computer Simulations. *Journal of Chemical Physics* **2003**, *119* (24), 12718–12728. https://doi.org/10.1063/1.1628670.

(6) Milano, G.; Müller-Plathe, F. Mapping Atomistic Simulations to Mesoscopic Models: A Systematic Coarse-Graining Procedure for Vinyl Polymer Chains. *Journal of Physical Chemistry B* **2005**, *109* (39), 18609–18619. https://doi.org/10.1021/jp0523571.

(7) Addison, C. I.; Hansen, J. P.; Krakoviack, V.; Louis, A. A. Coarse-Graining Diblock Copolymer Solutions: A Macromolecular Version of the Widom-Rowlinson Model. *Molecular Physics* **2005**, *103* (21–23), 3045–3054. https://doi.org/10.1080/00268970500186086.

(8) Sliozberg, Y. R.; Kröger, M.; Chantawansri, T. L. Fast Equilibration Protocol for Million Atom Systems of Highly Entangled Linear Polyethylene Chains. *Journal of Chemical Physics* **2016**, *144* (15), 154901. https://doi.org/10.1063/1.4946802.

(9) Müller-Plathe, F. Coarse-Graining in Polymer Simulation: From the Atomistic to the Mesoscopic Scale and Back. *ChemPhysChem* **2002**, *3*, 754–769. https://doi.org/1439-4235/02/03/09.

(10) Curro, J. G.; Schweizer, K. S. Theory of Polymer Melts: An Integral Equation Approach. *Macromolecules* **1987**, *20* (8), 1928–1934. https://doi.org/10.1021/ma00174a040.

(11) Chandler, D.; Andersen, H. C. Optimized Cluster Expansions for Classical Fluids. II. Theory of Molecular Liquids. *The Journal of Chemical Physics* **1972**, *57* (5), 1930–1937. https://doi.org/10.1063/1.1678513.

(12) Yatsenko, G.; Sambriski, E. J.; Nemirovskaya, M. A.; Guenza, M. Analytical Soft-Core Potentials for Macromolecular Fluids and Mixtures. *Physical Review Letters* **2004**, *93* (25), 1–4. https://doi.org/10.1103/PhysRevLett.93.257803.

(13) Clark, A. J.; Guenza, M. G. Mapping of Polymer Melts onto Liquids of Soft-Colloidal Chains. *Journal of Chemical Physics* **2010**, *132* (4), 044902. https://doi.org/10.1063/1.3292013.

(14) Clark, A. J.; McCarty, J.; Guenza, M. G. Effective Potentials for Representing Polymers in Melts as Chains of Interacting Soft Particles. *Journal of Chemical Physics* **2013**, *139* (12), 124906. https://doi.org/10.1063/1.4821818.

(15) Dinpajooh, M.; Guenza, M. G. On the Density Dependence of the Integral Equation Coarse-Graining Effective Potential. *Journal of Physical Chemistry B* **2018**, *122* (13), 3426–3440. https://doi.org/10.1021/acs.jpcb.7b10494.

(16) Ruhle, V.; Junghans, C.; Lukyanov, A.; Kremer, K.; Andrienko, D. Versatile Object-Oriented Toolkit for Coarse-Graining Applications. *Journal of Chemical Theory and Computation* **2009**, *5* (12), 3211–3223. https://doi.org/10.1021/ct900369w.

(17) Dinpajooh, M.; Guenza, M. G. Composition and Resolution Dependence of Effective Coarse-Graining Potentials in Multi-Resolution Simulations. *Soft Condensed Matter* **2019**. https://doi.org/arXiv:1906.01758v1.

(18) Dinpajooh, M.; Guenza, M. G. Coarse-Graining Simulation Approaches for Polymer Melts: The Effect of Potential Range on Computational Efficiency. *Soft Matter* **2018**, *14* (35), 7126–7144. https://doi.org/10.1039/c8sm00868j.

(19) Guenza, M. G.; Dinpajooh, M.; McCarty, J.; Lyubimov, I. Y. Accuracy, Transferability, and Efficiency of Coarse-Grained Models of Molecular Liquids. *Journal of Physical Chemistry B* **2018**, *122* (45), 10257–10278. https://doi.org/10.1021/acs.jpcb.8b06687.

(20) McCarty, J.; Clark, A. J.; Lyubimov, I. Y.; Guenza, M. G. Thermodynamic Consistency between Analytic Integral Equation Theory and Coarse-Grained Molecular Dynamics Simulations of Homopolymer Melts. *Macromolecules* **2012**, *45* (20), 8482–8493. https://doi.org/10.1021/ma301502w.

(21) Dinpajooh, M.; Guenza, M. G. Thermodynamic Consistency in the Structure-Based Integral Equation Coarse-Grained Method. *Polymer* **2017**, *117*, 282–286. https://doi.org/10.1016/j.polymer.2017.04.025.

(22) Chandler, D.; Andersen, H. C. Optimized Cluster Expansions for Classical Fluids. II. Theory of Molecular Liquids. *The Journal of Chemical Physics* **1972**, *57* (5), 1918–1929. https://doi.org/10.1063/1.1678512.

(23) Ghanbari, A.; Böhm, M. C.; Müller-Plathe, F. A Simple Reverse Mapping Procedure for Coarse-Grained Polymer Models with Rigid Side Groups. *Macromolecules* **2011**, *44* (13), 5520–5526. https://doi.org/10.1021/ma2005958.

(24) Ohkuma, T.; Kremer, K.; Daoulas, K. Equilibrating High-Molecular-Weight Symmetric and Miscible Polymer Blends with Hierarchical Back-Mapping. *Journal of Physics Condensed Matter* **2018**, *30* (17). https://doi.org/10.1088/1361-648X/aab684.

(25) Zhang, G.; Chazirakis, A.; Harmandaris, V. A.; Stuehn, T.; Daoulas, K. C.; Kremer, K. Hierarchical Modelling of Polystyrene Melts: From Soft Blobs to Atomistic Resolution. *Soft Matter* **2019**, *15* (2), 289–302. https://doi.org/10.1039/c8sm01830h.

(26) Nosé, S. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *The Journal of Chemical Physics* **1984**, *81* (1), 511–519. https://doi.org/10.1063/1.447334.

(27) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *Journal of Chemical Physics* **2007**, *126* (1), 014101. https://doi.org/10.1063/1.2408420.